



LREC 2026

**The Third Workshop on Patient-Oriented Language
Processing (CL4Health) @ LREC 2026**

Workshop Proceedings

Editors

**Deepak Gupta, Paul Thompson, Sophia Ananiadou and
Dina Demner-Fushman**

12 May 2026

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-
NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-51-7

Preface

Introduction

We are pleased to present the proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), held in conjunction with LREC 2026. This workshop continues to provide a dedicated forum for research on the spectrum of patient-oriented language processing and multimodal research. Following the success of the previous editions held at LREC-COLING 2024 and NAACL 2025, CL4Health has further established itself as a dedicated venue for advancing language technologies that support public health, including the summarization of health records for patient use, answering consumer health questions using reliable sources, detecting misinformation or potentially harmful content, and providing multimodal response when it more effectively supports patient understanding.

Such a venue is needed both to invigorate patient-oriented language processing research and to build a community of researchers interested in this area. The growing interest in this topic is fueled by several current trends:

- Rapid use of large language models for health counseling, diagnostic reasoning, and patient-facing question answering.
- Focus on simplifying and adapting clinical information through summarization, rewriting, and persona-based generation.
- Growth of multimodal tools for patient journeys, care pathways, and communication support.
- Increasing emphasis on safety, factual grounding, and standardized evaluation of patient-oriented systems.

The program includes contributions covering a broad spectrum of topics, including question answering, patient-oriented lay summarization, medical text simplification, and multimodal learning in the biomedical and consumer health domains. In addition to regular research papers, the workshop features four shared tasks, designed to encourage standardized evaluation and facilitate comparative analysis across systems. These shared tasks highlight current progress and persistent challenges in developing robust and trustworthy patient-oriented language technologies.

Invited Speaker

Kailai Yang, University of Manchester, UK

Towards Trustworthy and Efficient Continual Learning for Large Language Models

Current Large Language Models (LLMs) continue to face several key limitations in adapting to specific domains. First, their ability to model human values and learn rich representations remains limited. Second, general-purpose LLMs often lack sufficient domain-specific

knowledge and alignment for professional applications. Third, existing continual learning methods are still prone to catastrophic forgetting, instability in multi-objective optimization, and limited training efficiency. This research addresses these challenges from three complementary directions. First, we improve representation learning for language models through knowledge fusion and innovative paradigms, including contrastive and disentangled learning methods. Second, we strengthen domain alignment via domain knowledge injection, in-context learning, and advanced post-training approaches. Third, we improve continual learning for LLMs by proposing novel automated data-mixing strategies, selective preference optimization method for efficient training, and a plug-and-play multi-objective alignment framework that generalizes across models and objectives. Overall, this research provides practical and effective solutions to enhance the reliability and efficiency for domain adaptation for LLMs.

Brief Biography

Kailai Yang is an incoming researcher at the University of Manchester. His main research areas are artificial intelligence and large language models. He has authored over 20 papers with over 2,000 citations in total. His research as leading contributor has been published in leading conferences and journals, such as NeurIPS, ACL, EMNLP, WWW, and KDD. His work has also been featured by renowned media such as Forbes.

Shared Tasks

- **Detecting Dosing Errors from Clinical Trials (CT-DEB'26).**

The Clinical Trials Dosing Errors Benchmark 2026 (CT-DEB'26) is dedicated to automated detection of the risks of medication dosing errors within clinical trial protocols. Leveraging a curated dataset of over 29K trial records from the ClinicalTrials.gov registry, participants are challenged to predict the risk probabilities of protocols likely to result in dosing errors.

- **Automatic Case Report Form (CRF) Filling from Clinical Notes.**

Case Report Forms (CRFs) are standardized instruments in medical research that collect patient data in a consistent and reliable manner. They consist of a predefined list of items to be filled with patient information. Each item aims to collect a portion of the information relevant to a specific clinical goal (e.g., allergies, disease chronicity, test results). Automating CRF filling from clinical notes would accelerate clinical research, reduce manual burden on healthcare professionals, and create structured representations that can be directly leveraged to produce accessible, patient- and practitioner-friendly summaries.

- **ArchEHR-QA 2026: Grounded Question Answering from Electronic Health Records.**

ArchEHR-QA targets the problem of producing answers to patient questions that are supported by and explicitly linked to the underlying clinical notes. This second iteration builds on the 2025 challenge (co-located with the ACL 2025 BioNLP Workshop) by expanding the dataset and introducing four complementary subtasks covering question interpretation, clinical evidence identification, answer generation, and answer–evidence alignment.

- **FoodBench-QA 2026: Grounded Food & Nutrition Question Answering.**

FoodBench-QA 2026 is a shared task challenging systems to answer food and nutrition questions using evidence from nutrient databases and food ontologies. The dataset

includes realistic dietary queries, ingredient and their quantities lists, and recipe descriptions, requiring models to perform nutrient estimation, FSA traffic-light prediction, and food entity recognition/linking across three food semantic models.

Submissions

CL4Health received 36 valid submissions for the main workshop, of which 11 were accepted as oral presentations, 17 as posters, and 8 were rejected. The work covers a wide range of topics focusing on patients' well-being and proper care. The topics include question answering, medical imaging, patient-oriented lay summarization, medical text simplification, mental health issues, and communication support. We received 28 valid submissions for the shared task papers, of which 25 were accepted as posters and 3 were rejected.

We would like to thank all the authors for their submissions. We are grateful to the shared task organizers for their efforts in designing and coordinating the evaluation benchmarks. Our sincere thanks go to the invited speaker and participants for their engagement and contributions to the workshop. Finally, we acknowledge the Program Committee (listed elsewhere in this volume) for their dedicated efforts in evaluating the submissions within a short review period and identifying work ready for presentation, as well as studies that would benefit from further experimentation and analysis informed by reviewer comments. We hope that the work presented in these proceedings will contribute to ongoing research in patient-centered language technologies and foster further advances in accessible, reliable, and effective language technologies for public health.

Deepak Gupta, Paul Thompson, Sophia Ananiadou and Dina Demner-Fushman (Organizers).

Organizing Committee

Deepak Gupta, National Library of Medicine, USA
Paul Thompson, University of Manchester, UK
Sophia Ananiadou, University of Manchester, UK & Archimedes/Athena Research Center, Greece
Dina Demner-Fushman, National Library of Medicine, USA

Shared Task Organizers

Detecting Dosing Errors from Clinical Trials (CT-DEB'26)

Sohrab Ferdowsi, University of Geneva, Switzerland
Félicien Hêche, University of Geneva, Switzerland
Anthony Yazdani, University of Geneva, Switzerland
Edward Choi, KAIST, Korea
Douglas Teodoro, University of Geneva, Switzerland

Automatic Case Report Form (CRF) Filling from Clinical Notes

Pietro Ferrazzi, Fondazione Bruno Kessler, Trento, Italy
Soumitra Ghosh, Fondazione Bruno Kessler, Trento, Italy
Alberto Lavelli, Fondazione Bruno Kessler, Trento, Italy
Bernardo Magnini, Fondazione Bruno Kessler, Trento, Italy

ArchEHR-QA 2026: Grounded Question Answering from Electronic Health Records

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

FoodBench-QA 2026: Grounded Food & Nutrition Question Answering

Tome Eftimov, Jožef Stefan Institute, Ljubljana, Slovenia
Ana Gjorgjevikj, Jožef Stefan Institute, Ljubljana, Slovenia
Matej Martinc, Jožef Stefan Institute, Ljubljana, Slovenia
Gjorgjina Cenikj, Jožef Stefan Institute, Ljubljana, Slovenia
Sašo Džeroski, Jožef Stefan Institute, Ljubljana, Slovenia
Barbara Koroušić Seljak, Jožef Stefan Institute, Ljubljana, Slovenia

Program Committee

Sophia Ananiadou, University of Manchester, UK & Archimedes/Athena Research Center, Greece
Judith Jeyafreeda Andrew, Université Paris Cité, Paris, France
Mohammad Arvan, University of Illinois Chicago, USA
Zainab Awan, Queen Mary University of London, UK
Abby Blocker, University of Cape Town, South Africa
Giovanni Bonetta, Fondazione Bruno Kessler, Italy
Leonardo Campillos-Llanos, Spanish National Research Council, Spain
Gjorgjina Cenikj, Jožef Stefan Institute, Slovenia
Surabhi Datta, IMO Health, USA
Dina Demner-Fushman, National Library of Medicine, USA
Luise Duerlich, Research Institutes of Sweden, Sweden
Tome Eftimov, Jožef Stefan Institute, Slovenia

Sohrab Ferdowsi, University of Geneva, Switzerland
Pietro Ferrazzi, Fondazione Bruno Kessler, Italy
Manas Gaur, University of Maryland Baltimore County, USA
Soumitra Ghosh, Fondazione Bruno Kessler, Italy
Natalia Grabar, Université de Lille, France
Tudor Groza, Bioinformatics Institute, A*STAR, Singapore
Deepak Gupta, National Library of Medicine, USA
Félicien Hêche, University of Geneva, Switzerland
Sy Hwang, University of Pennsylvania, USA
Samah Jarad-Fodeh, Yale School of Medicine, USA
Dimitrios Kokkinakis, University of Gothenberg, Sweden
Saran Krishnasamy, Gigit.ai, USA
Tiziano Labruna, University of Padova, Italy
Vojtech Lanz, Charles University, Czech Republic
Alberto Lavelli, Fondazione Bruno Kessler, Italy
Chuyuan Li, University of British Columbia, Canada
Matej Martinc, Jožef Stefan Institute, Slovenia
Vineet Mishra, HCLTech, USA
Anindita Nath, University of Texas at El Paso, USA
Aleksandr Nesterov, Artificial Intelligence Research Institute, USA
Aurélié Névél, Université de Paris-Saclay, CNRS, LISN, Orsay, France
Brian Ondov, Yale School of Medicine, USA
Tabea Pakull, University Hospital Essen, Germany
Mehdi Parviz, University of Copenhagen, Denmark
Priyanshu Priya, Indian Institute of Technology Patna, India
Roland Roller, German Research Center for Artificial Intelligence (DFKI), Germany
Roshan Santhosh, Adobe, USA
Abeed Sarker, Emory University, USA
Apoorva Singh, Fondazione Bruno Kessler, Italy
Sarvesh Soni, National Library of Medicine, USA
Suzanne Tamang, Stanford University, USA
Douglas Teodoro, Fondazione Bruno Kessler, Italy
Paul Thompson, University of Manchester, UK
Talia Tseriotou, Queen Mary University of London, UK
Grigorios Tsoumakas, Aristotle University of Thessaloniki, Greece
Shubham Vatsal, New York University, USA
Jinghua Xu, Heidelberg University, Germany
Zhicheng Yang, PAII Inc., USA
Anthony Yazdani, University of Geneva, Switzerland
Tianlin Zhang, CHN Energy, China
Pierre Zweigenbaum, LISN, CNRS, Université Paris-Saclay, France

Table of Contents

<i>FHIRPath-QA: Executable Question Answering over FHIR Electronic Health Records</i>	
Michael Frew, Nishit Bheda and Bryan Tripp	1
<i>COACH Meets QUORUM: A Framework and Pipeline for Aligning User, Expert, and Developer Perspectives in LLM-Generated Health Counselling</i>	
Yee Man Ng, Bram van Dijk, Pieter Beynen, Otto Boekesteijn, Joris Jansen, Gerard van Oortmerssen, Max J. van Duijn and Marco Spruit	15
<i>Addressing Domain Shift in Health Coaching Note Analysis through Factorized Synthetic Data Generation</i>	
Michael Tänzer, Iva Bojic, Ashwini Yuvraj Lawate, Andy Hau Yan Ho and Andy Khong .	26
<i>Scalable Generation of Adult-Oriented Therapeutic Reading Texts for Russian Aphasia Rehabilitation</i>	
Anastasia Kolmogorova, Anastasia Margolina, Alina Telnova and Igor Ilchenko	41
<i>An Open-Resource Knowledge Augmentation for Biomedical Lay Summarization</i>	
João Pedro Veloso and Evelin Amorim	50
<i>TabMedQA: From Structured Data to Question-Answer Datasets in Early Clinical Decision-Making</i>	
Gabriel Iturra Bocaz, Petra Galuščáková, Sol Gedde Vedde and Alvaro Fernandez-Quilez	61
<i>SimpliMED: Automatic Simplification of Cardiology Discharge Reports Using Large Language Models</i>	
Lucas Molino-Piñar, Manuel Carlos Díaz Galiano, María-Teresa Martín-Valdivia, Jose Angel Urbano-Moral and Elena Sola-Garcia	72
<i>Dutch Metaphor Extraction from Cancer Patients' Interviews and Forum Data Using LLMs and Human in the Loop</i>	
Lifeng Han, David Lindevelt, Sander Puts, Erik van Mulligen and Suzan Verberne	82
<i>Compressed Representations of Patient Records: A Comparative Study of Template-Based and LLM-Based Methods for Clinical Data Summarization and Visualization</i>	
Andreas Stöckl, Oliver Krauss and Sophie Bauernfeind	96
<i>Medical Text Rewriting for Non-Experts: A Guideline-Driven LLM Approach</i>	
Mana Kuramoto, Hiroyuki Nagai, Keiko Yamada, Hiroo Ide, Masayo Hayakawa, Tomohiro Nishiyama, Shoko Wakamiya and Eiji Aramaki	107
<i>Patient-Specific Care Pathway Visualisation for Medical Nursing Staff</i>	
Sophie Bauernfeind, Selina Adlberger, Oliver Krauss and Andreas Stöckl	117
<i>What Makes a Good Doctor Response? A Study on Text-Based Telemedicine</i>	
Adrian Cosma, Cosmin Dumitrache and Emilian Radoi	132
<i>Datasets for a Chatbot for Clinical Trial Search</i>	
Yumeng Yang, Ethan Ludmir and Kirk Roberts	139

<i>HealthTrajectory: Patient Journey Summaries and Visualizations for Patient-Clinician Communication Support</i>	
Rohmah Hidayah, Tomohiro Nishiyama, Shoko Wakamiya and Eiji Aramaki	149
<i>Medical-FLAVORS-AECC: Spanish Oncological Metaphors Dataset</i>	
Lucia Pitarch, Jordi Bernad, Sergio-Luis Ojeda-Trueba, Alec Sánchez-Montero, Maxim Ionov, Emma Anglés-Herrero, Ángel Óscar Corona Beomont and Gemma Bel-Enguix	159
<i>A Synthetic Conversational Dataset for Type 2 Diabetes Management</i>	
Stergios Ntanavaras, Maaïke de Boer and Piek T.J.M. Vossen	171
<i>Italian Medical Term Simplification: From Patient Information Leaflets to Simplified Language Resources</i>	
Maria Pia di Buono	182
<i>Evaluating Professional Acceptability of LLM-Generated Systematic Review Summaries in Healthcare: Psychiatrists' Perspectives</i>	
Paul Thompson, Artemis Boulogeorgou, Fotini Kaponi, Efstathia Soufleri and Sophia Ananiadou	190
<i>Reasoning, Contrastive, and In-Context Strategies for Opioid Use Stage Detection on Social Media</i>	
Vinu Ekanayake and Ramakanth Kavuluru	207
<i>FoodBench-QA: Overview of the Shared Task on Grounded Food and Nutrition Question Answering</i>	
Tome Eftimov, Ana Gjorgjevikj, Matej Martinc, Gjorgjina Cenikj, Sašo Džeroski and Barbara Koroušič Seljak	223
<i>Overview of the CT-DEB'26 Shared Task on Predicting Dosing Errors in Interventional Clinical Trials</i>	
Sohrab Ferdowsi, Félicien Hêche, Anthony Yazdani, Edward Choi, Sara Sansaloni-Pastor and Douglas Teodoro	234
<i>Overview of the CRF 2026 Shared Task on Clinical Case Report Forms Filling</i>	
Pietro Ferrazzi, Soumitra Ghosh, Alberto Lavelli and Bernardo Magnini	245
<i>Overview of the ArchEHR-QA 2026 Shared Task on Grounded Question Answering from Electronic Health Records</i>	
Sarvesh Soni and Dina Demner-Fushman	255
<i>Structured Radiology Intelligence: Extracting Structured Data from MRI Reports Using LLMs</i>	
Sushvin Marimuthu, Parameswari Krishnamurthy, Dipti Misra Sharma, Goldwin H, Anu Eapen, Betty Simon and Anuradha Chandramohan	268
<i>Useful to Whom? A Persona-Driven Evaluation of Knowledge-Adapted Health Question Reformulation via LLM Simulation</i>	
Jooyeon Lee, Luan Huy Pham and Özlem Uzuner	281
<i>MedGore: An Approach and a Dataset for Identification of Sensitive Medical Images</i>	
Soumya Gayen, Rory Mulcahey, Russell Loane, Dina Demner-Fushman and Deepak Gupta	296

<i>Diagnostic Reasoning with Large Language Models for a Rare Disease: Case Study of Primary Ciliary Dyskinesia</i>	
Swati Rajwal, Mary Ellen M. Fain, Lokesh Guglani and Abeed Sarker	307
<i>Beyond One-Size-Fits-All: Multi-Agent Refinement Framework for Persona-Based Biomedical Summarization</i>	
Rohan Charudatt Salvi, Chirag Chawla, Md. Shad Akhtar and Shweta Yadav	320
<i>Automated Detection of Dosing Errors in Clinical Trial Narratives: A Multi-Modal Feature Engineering Approach with LightGBM</i>	
Mohammad AL-Smadi	333
<i>CaresAI at CT-DEB'26: Detecting Dosing Errors In Clinical Trials Using Domain-Specific Transformer Embeddings and Classification Models</i>	
Leon Hamnett, Favour Igwezeke, Joseph Itopa Abubakar and Mary Adetutu Adewunmi	344
<i>CGU-ILALab at FoodBench-QA 2026: Comparing Traditional and LLM-based Approaches for Recipe Nutrient Estimation</i>	
Wei-Chun Chen, Yu-Xuan Chen, I-Fang Chung and Ying-Jia Lin	362
<i>FCProfiler: Structured and Deterministic Pipeline for Recipe-Level Nutrient Estimation</i>	
Siyoon Kim	369
<i>DocUA at CRF Filling 2026: LLM StructCore — Schema-Guided Reasoning Condensation and Deterministic Compilation</i>	
Serhii Zabolotnii	375
<i>GREYC at CRF Filling 2026: Rewrite Before You Extract - Rewriting Clinical Notes for Automated CRF</i>	
Jesus Lovon-Melgarejo, J������ Pantin and Ga���� Dias	384
<i>Innov8rs at CRF Filling 2026: An Iterative Multi-LLM Ensemble Pipeline with Dynamic Few-Shot Retrieval and Data-Driven Precision Filtering</i>	
Samminga Sainath Rao, Sumit Mishra and Chanchal Suman	390
<i>Aurum at CRF Filling 2026: Modular DSPy Extractors with Qwen3-Max for Multilingual CRF Filling</i>	
Vinay Babu Ulli, Jyoti Kumari and Anindita Mondal	395
<i>sebis at CRF Filling 2026: A Two-Stage Local LLM Pipeline for Medical CRF Filling</i>	
Katharina Sommer, Tristan Till and Florian Matthes	402
<i>Polimi at CRF Filling 2026: Prompt-Based Information Extraction from Italian Clinical Notes</i>	
Vittorio Torri and Francesca Ieva	412
<i>Cohere Labs Community at FoodBench-QA 2026: The Cake Makes the Ingredients</i>	
Ravi Ranjan, Roshan Santhosh and Lucien Carroll	428
<i>HiTZ-IXA at ArchEHR-QA 2026: Evidence Alignment Through Self-Consistency and Prompt Curation in Memory-Constrained Environments</i>	
Xabier Irastortza-Urbieta, Maite Oronoz and Alicia P������	434

<i>BIT.UA-AAUBS at ArchEHR-QA 2026: Evaluating Open-Source and Proprietary LLMs via Prompting in Low-Resource QA</i>	
Richard A. A. Jonker, Alexander Christiansen, Alexandros Maniatis, Rúben Garrido, Rogério Braunschweiger de Freitas Lima, Roman Jurowetzki and Sérgio Matos	441
<i>WisPerMed at ArchEHR-QA 2026: Retrieval-Augmented Prompting for Grounded EHR Question Answering</i>	
Jan-Henning Büns, Tabea Margareta Grace Pakull, Hendrik Damm, Bohao Chu, Christoph M. Friedrich, Felix Nensa, Elisabeth Livingstone, Peter A. Horn and Norbert Fuhr	455
<i>sebis at ArchEHR-QA 2026: How Much Can You Do Locally? Evaluating Grounded EHR QA on a Single Notebook</i>	
Ibrahim Ebrar Yurt, Fabian Tobias Karl, Tejaswi Chopra and Florian Matthes	469
<i>MedEvi-NS at ArchEHR-QA 2026: Using Clinical Reasoning Principles to Improve Zero-shot Capabilities of Large Language Models in Evidence Alignment</i>	
Mengxuan Sun and Nicolay Rusnachenko	482
<i>UIC-AIHealth4All at ArchEHR-QA 2026: Answer-First Evidence Grounding for Clinical Question Answering</i>	
Mohammad Arvan, Hossein Haeri, Natalie Parde and Rebecca Feinstein	487
<i>tt501 at ArchEHR-QA 2026: Few-Shot Prompting with Retrieval-Augmented Generation for Grounded Clinical EHR Question Answering</i>	
Tai Tan Tran	497
<i>HealthNLP_Retrievers at ArchEHR-QA 2026: Cascaded LLM Pipeline for Grounded Clinical Question Answering</i>	
Md Biplob Hosen, Md Alomgeer Hussein, Md Akmol Masud, Omar Faruque, Tera L Reynolds and Lujie Karen Chen	506
<i>Yale-DM-Lab at ArchEHR-QA 2026: Deterministic Grounding and Multi-Pass Evidence Alignment for EHR Question Answering</i>	
Elyas Irankhah and Samah Fodeh	515
<i>Razreshili at ArchEHR-QA 2026: Evidence Alignment via LLM Prompting and Cross-Encoder Fine-tuning</i>	
Arina Zemchyk	524
<i>Neural at ArchEHR-QA 2026: One Method Fits All: Unified Prompt Optimization for Clinical QA over EHRs</i>	
Abrar Majeedi, Viswanatha Reddy Gajjala, Sai Prasanna Teja Reddy Bogireddy and Siddhant Rai	530
<i>OptiMed at ArchEHR-QA 2026: GEPA Prompt Optimization and Multi-Agent Majority Voting for EHR-Grounded Question Answering</i>	
Feras AlMannaa, Talia Tseriotou and Maria Liakata	539
<i>TAMU-NLP at ArchEHR-QA 2026: Grounded Clinical QA with Evidence Identification and Intent-Aware Answer Generation</i>	
Xinqi Su, Rongrong Wang, Sunyang Fu, Hongfang Liu and Ruihong Huang	551

GigitAI at ArchEHR-QA 2026: Prompting Strategies and Constitutional AI for Clinical Question Answering

Saran Krishnasamy and Inez Wihardjo	565
---	-----

Workshop Program

Tuesday, May 12, 2026

09:00–09:10 Opening remarks

09:10–10:10 Session 1: Oral Presentations

09:10–09:30 *FHIRPath-QA: Executable Question Answering over FHIR Electronic Health Records*

Michael Frew, Nishit Bheda and Bryan Tripp

09:30–09:50 *COACH Meets QUORUM: A Framework and Pipeline for Aligning User, Expert, and Developer Perspectives in LLM-Generated Health Counselling*

Yee Man Ng, Bram van Dijk, Pieter Beynen, Otto Boekesteijn, Joris Jansen, Gerard van Oortmerssen, Max J. van Duijn and Marco Spruit

09:50–10:10 *Addressing Domain Shift in Health Coaching Note Analysis through Factorized Synthetic Data Generation*

Michael Tänzer, Iva Bojic, Ashwini Yuvraj Lawate, Andy Hau Yan Ho and Andy Khong

10:10–10:31 Session 2: Virtual Poster Presentations

10:10–10:17 *Scalable Generation of Adult-Oriented Therapeutic Reading Texts for Russian Aphasia Rehabilitation*

Anastasia Kolmogorova, Anastasia Margolina, Alina Telnova and Igor Ilchenko

10:17–10:24 *An Open-Resource Knowledge Augmentation for Biomedical Lay Summarization*

João Pedro Veloso and Evelin Amorim

10:24–10:31 *TabMedQA: From Structured Data to Question-Answer Datasets in Early Clinical Decision-Making*

Gabriel Iturra Bocaz, Petra Galuščáková, Sol Gedde Vedde and Alvaro Fernandez-Quilez

10:31–11:00 Coffee Break and Poster Session

SimpliMED: Automatic Simplification of Cardiology Discharge Reports Using Large Language Models

Lucas Molino-Piñar, Manuel Carlos Díaz Galiano, María-Teresa Martín-Valdivia, Jose Angel Urbano-Moral and Elena Sola-Garcia

Tuesday, May 12, 2026 (continued)

Dutch Metaphor Extraction from Cancer Patients' Interviews and Forum Data Using LLMs and Human in the Loop

Lifeng Han, David Lindevelt, Sander Puts, Erik van Mulligen and Suzan Verberne

Compressed Representations of Patient Records: A Comparative Study of Template-Based and LLM-Based Methods for Clinical Data Summarization and Visualization

Andreas Stöckl, Oliver Krauss and Sophie Bauernfeind

Medical Text Rewriting for Non-Experts: A Guideline-Driven LLM Approach

Mana Kuramoto, Hiroyuki Nagai, Keiko Yamada, Hiroo Ide, Masayo Hayakawa, Tomohiro Nishiyama, Shoko Wakamiya and Eiji Aramaki

Patient-Specific Care Pathway Visualisation for Medical Nursing Staff

Sophie Bauernfeind, Selina Adlberger, Oliver Krauss and Andreas Stöckl

What Makes a Good Doctor Response? A Study on Text-Based Telemedicine

Adrian Cosma, Cosmin Dumitrache and Emilian Radoi

Datasets for a Chatbot for Clinical Trial Search

Yumeng Yang, Ethan Ludmir and Kirk Roberts

HealthTrajectory: Patient Journey Summaries and Visualizations for Patient-Clinician Communication Support

Rohmah Hidayah, Tomohiro Nishiyama, Shoko Wakamiya and Eiji Aramaki

Medical-FLAVORS-AECC: Spanish Oncological Metaphors Dataset

Lucia Pitarch, Jordi Bernad, Sergio-Luis Ojeda-Trueba, Alec Sánchez-Montero, Maxim Ionov, Emma Anglés-Herrero, Ángel Óscar Corona Beomont and Gemma Bel-Enguix

A Synthetic Conversational Dataset for Type 2 Diabetes Management

Stergios Ntanavaras, Maaïke de Boer and Piek T.J.M. Vossen

Italian Medical Term Simplification: From Patient Information Leaflets to Simplified Language Resources

Maria Pia di Buono

Tuesday, May 12, 2026 (continued)

11:00–11:40 Session 3: Oral Presentations

11:00–11:20 *Evaluating Professional Acceptability of LLM-Generated Systematic Review Summaries in Healthcare: Psychiatrists' Perspectives*
Paul Thompson, Artemis Boulogeorgou, Fotini Kaponi, Efstathia Soufleri and Sophia Ananiadou

11:20–11:40 *Reasoning, Contrastive, and In-Context Strategies for Opioid Use Stage Detection on Social Media*
Vinu Ekanayake and Ramakanth Kavuluru

11:40–12:40 Session 4: Shared Task Overviews

11:40–11:55 *FoodBench-QA: Overview of the Shared Task on Grounded Food and Nutrition Question Answering*
Tome Eftimov, Ana Gjorgjevikj, Matej Martinc, Gjorgjina Cenikj, Sašo Džeroski and Barbara Koroušič Seljak

11:55–12:10 *Overview of the CT-DEB'26 Shared Task on Predicting Dosing Errors in Interventional Clinical Trials*
Sohrab Ferdowsi, Félicien Hêche, Anthony Yazdani, Edward Choi, Sara Sansaloni-Pastor and Douglas Teodoro

12:10–12:25 *Overview of the CRF 2026 Shared Task on Clinical Case Report Forms Filling*
Pietro Ferrazzi, Soumitra Ghosh, Alberto Lavelli and Bernardo Magnini

12:25–12:40 *Overview of the ArchEHR-QA 2026 Shared Task on Grounded Question Answering from Electronic Health Records*
Sarvesh Soni and Dina Demner-Fushman

Tuesday, May 12, 2026 (continued)

12:40–13:01 Session 5: Virtual Poster Presentations

12:40–12:47 *Structured Radiology Intelligence: Extracting Structured Data from MRI Reports Using LLMs*
Sushvin Marimuthu, Parameswari Krishnamurthy, Dipti Misra Sharma, Goldwin H, Anu Eapen, Betty Simon and Anuradha Chandramohan

12:47–12:54 *Useful to Whom? A Persona-Driven Evaluation of Knowledge-Adapted Health Question Reformulation via LLM Simulation*
Jooyeon Lee, Luan Huy Pham and Özlem Uzuner

12:54–13:01 *MedGore: An Approach and a Dataset for Identification of Sensitive Medical Images*
Soumya Gayen, Rory Mulcahey, Russell Loane, Dina Demner-Fushman and Deepak Gupta

13:01–14:00 Lunch Break

14:00–14:35 Session 6: Oral Presentations

14:00–14:20 *Diagnostic Reasoning with Large Language Models for a Rare Disease: Case Study of Primary Ciliary Dyskinesia*
Swati Rajwal, Mary Ellen M. Fain, Lokesh Guglani and Abeed Sarker

14:20–14:35 *Beyond One-Size-Fits-All: Multi-Agent Refinement Framework for Persona-Based Biomedical Summarization*
Rohan Charudatt Salvi, Chirag Chawla, Md. Shad Akhtar and Shweta Yadav

14:35–15:05 *Invited Talk: Towards Trustworthy and Efficient Continual Learning for Large Language Models*
Kailai Yang

Tuesday, May 12, 2026 (continued)

15:05–16:01 Session 7: Shared Task Virtual Poster Presentations

15:05–15:12 *Automated Detection of Dosing Errors in Clinical Trial Narratives: A Multi-Modal Feature Engineering Approach with LightGBM*
Mohammad AL-Smadi

15:12–15:19 *CaresAI at CT-DEB'26: Detecting Dosing Errors In Clinical Trials Using Domain-Specific Transformer Embeddings and Classification Models*
Leon Hamnett, Favour Igwezeke, Joseph Itopa Abubakar and Mary Adetutu Adewunmi

15:19–15:26 *CGU-ILALab at FoodBench-QA 2026: Comparing Traditional and LLM-based Approaches for Recipe Nutrient Estimation*
Wei-Chun Chen, Yu-Xuan Chen, I-Fang Chung and Ying-Jia Lin

15:26–15:33 *FCProfiler: Structured and Deterministic Pipeline for Recipe-Level Nutrient Estimation*
Siyoon Kim

15:33–15:40 *DocUA at CRF Filling 2026: LLM StructCore — Schema-Guided Reasoning Condensation and Deterministic Compilation*
Serhii Zabolotnii

15:40–15:47 *GREYC at CRF Filling 2026: Rewrite Before You Extract - Rewriting Clinical Notes for Automated CRF*
Jesus Lovon-Melgarejo, Jérémie Pantin and Gaël Dias

15:47–15:54 *Innov8rs at CRF Filling 2026: An Iterative Multi-LLM Ensemble Pipeline with Dynamic Few-Shot Retrieval and Data-Driven Precision Filtering*
Samminga Sainath Rao, Sumit Mishra and Chanchal Suman

15:54–16:01 *Aurum at CRF Filling 2026: Modular DSPy Extractors with Qwen3-Max for Multilingual CRF Filling*
Vinay Babu Ulli, Jyoti Kumari and Anindita Mondal

16:01–17:00 Coffee Break and Poster session

sebis at CRF Filling 2026: A Two-Stage Local LLM Pipeline for Medical CRF Filling
Katharina Sommer, Tristan Till and Florian Matthes

Polimi at CRF Filling 2026: Prompt-Based Information Extraction from Italian Clinical Notes
Vittorio Torri and Francesca Ieva

Tuesday, May 12, 2026 (continued)

Cohere Labs Community at FoodBench-QA 2026: The Cake Makes the Ingredients

Ravi Ranjan, Roshan Santhosh and Lucien Carroll

HiTZ-IXA at ArchEHR-QA 2026: Evidence Alignment Through Self-Consistency and Prompt Curation in Memory-Constrained Environments

Xabier Irastortza-Urbieta, Maite Oronoz and Alicia Pérez

BIT.UA-AAUBS at ArchEHR-QA 2026: Evaluating Open-Source and Proprietary LLMs via Prompting in Low-Resource QA

Richard A. A. Jonker, Alexander Christiansen, Alexandros Maniatis, Rúben Garrido, Rogério Braunschweiger de Freitas Lima, Roman Jurawetzki and Sérgio Matos

WisPerMed at ArchEHR-QA 2026: Retrieval-Augmented Prompting for Grounded EHR Question Answering

Jan-Henning Büns, Tabea Margareta Grace Pakull, Hendrik Damm, Bohao Chu, Christoph M. Friedrich, Felix Nensa, Elisabeth Livingstone, Peter A. Horn and Norbert Fuhr

sebis at ArchEHR-QA 2026: How Much Can You Do Locally? Evaluating Grounded EHR QA on a Single Notebook

Ibrahim Ebrar Yurt, Fabian Tobias Karl, Tejaswi Chopra and Florian Matthes

MedEvi-NS at ArchEHR-QA 2026: Using Clinical Reasoning Principles to Improve Zero-shot Capabilities of Large Language Models in Evidence Alignment

Mengxuan Sun and Nicolay Rusnachenko

UIC-AIHealth4All at ArchEHR-QA 2026: Answer-First Evidence Grounding for Clinical Question Answering

Mohammad Arvan, Hossein Haeri, Natalie Parde and Rebecca Feinstein

17:00–17:56 Session 8: Shared Task Virtual Poster Presentations

17:00–17:07 *tt501 at ArchEHR-QA 2026: Few-Shot Prompting with Retrieval-Augmented Generation for Grounded Clinical EHR Question Answering*
Tai Tan Tran

17:07–17:14 *HealthNLP_Retrievers at ArchEHR-QA 2026: Cascaded LLM Pipeline for Grounded Clinical Question Answering*
Md Biplob Hosen, Md Alomgeer Hussein, Md Akmol Masud, Omar Faruque, Tera L Reynolds and Lujie Karen Chen

Tuesday, May 12, 2026 (continued)

- 17:14–17:21 *Yale-DM-Lab at ArchEHR-QA 2026: Deterministic Grounding and Multi-Pass Evidence Alignment for EHR Question Answering*
Elyas Irankhah and Samah Fodeh
- 17:21–17:28 *Razreshili at ArchEHR-QA 2026: Evidence Alignment via LLM Prompting and Cross-Encoder Fine-tuning*
Arina Zemchyk
- 17:28–17:35 *Neural at ArchEHR-QA 2026: One Method Fits All: Unified Prompt Optimization for Clinical QA over EHRs*
Abrar Majeedi, Viswanatha Reddy Gajjala, Sai Prasanna Teja Reddy Bogireddy and Siddhant Rai
- 17:35–17:42 *OptiMed at ArchEHR-QA 2026: GEPA Prompt Optimization and Multi-Agent Majority Voting for EHR-Grounded Question Answering*
Feras AlMannaa, Talia Tseriotou and Maria Liakata
- 17:42–17:49 *TAMU-NLP at ArchEHR-QA 2026: Grounded Clinical QA with Evidence Identification and Intent-Aware Answer Generation*
Xinqi Su, Rongrong Wang, Sunyang Fu, Hongfang Liu and Ruihong Huang
- 17:49–17:56 *GigitAI at ArchEHR-QA 2026: Prompting Strategies and Constitutional AI for Clinical Question Answering*
Saran Krishnasamy and Inez Wihardjo
- 17:56–18:00 Closing Remarks**

FHIRPath-QA: Executable Question Answering over FHIR Electronic Health Records

Michael Frew, Nishit Bheda, Bryan Tripp

University of British Columbia; University of Waterloo

michael.frew@ece.ubc.ca, nishit.bheda@uwaterloo.ca, bptripp@uwaterloo.ca

Abstract

Though patients are increasingly granted digital access to their electronic health records (EHRs), existing interfaces may not support precise, trustworthy answers to patient-specific questions. Large language models (LLM) show promise in clinical question answering (QA), but retrieval-based approaches are computationally inefficient, prone to hallucination, and difficult to deploy over real-life EHRs. This work introduces FHIRPath-QA, the first open dataset and benchmark for patient-specific QA that includes open-standard FHIRPath queries over real-world clinical data. A text-to-FHIRPath QA paradigm is proposed that shifts reasoning from free-text generation to FHIRPath query synthesis. For o4-mini, this reduced average token usage by $391\times$ relative to retrieval-first prompting (629,829 vs 1,609 tokens per question) and lowered failure rates from 0.36 to 0.09 on clinician-phrased questions. Built on MIMIC-IV on FHIR Demo, the dataset pairs over 14k natural language questions in patient and clinician phrasing with validated FHIRPath queries and answers. Empirically, the evaluated LLMs achieve at most 42% accuracy, highlighting the challenge of the task, but benefit strongly from supervised fine-tuning, with query synthesis accuracy improving from 27% to 79% for 4o-mini. These results highlight that text-to-FHIRPath synthesis has the potential to serve as a practical foundation for safe, efficient, and interoperable consumer health applications, and the FHIRPath-QA dataset and benchmark serve as a starting point for future research on the topic. The full dataset and generation code can be accessed on [GitHub](#).

Keywords: FHIRPath, Question-Answering, Electronic Health Records

1. Introduction

Over the past decade, healthcare systems worldwide have adopted standardized, interoperable representations of electronic health records (EHRs). HL7 Fast Healthcare Interoperability Resources (FHIR) has emerged as the dominant framework, with national adoption across major EHR vendors and mandated API requirements in jurisdictions such as the United States ([HL7 International; Organization et al., 2023](#)).

However, digital access does not equal understanding. Although patients can increasingly view their EHRs, FHIR's graph-structured design makes navigation difficult ([Zhang and Walji, 2014; Roman et al., 2017](#)). Most systems rely on keyword search, often returning entire documents or dense tables rather than the specific value requested ([Hanauer et al., 2015](#)). Bridging this gap requires automated question-answering (QA) systems capable of retrieving precise, verifiable answers from a patient's own record ([Del Fiore et al., 2014](#)).

Large Language Models (LLMs) demonstrate strong clinical reasoning on general medical QA tasks ([Bardhan et al., 2024](#)). However, deploying them over patient-specific EHR data introduces substantial challenges. Context windows are finite, making full-record ingestion costly and prone to degradation as length increases ([Liu et al., 2024; Chen et al., 2024](#)). Retrieval-Augmented Generation (RAG) systems remain vulnerable to hallu-

cination ([Zhang and Zhang, 2025](#)), which is unacceptable in patient-facing tools where incorrect answers may influence health decisions ([Bussone et al., 2015](#)). Additionally, LLMs struggle to reliably traverse FHIR's deeply nested graph structure; benchmarks such as FHIR-AgentBench report consistent failures in multi-hop reasoning and reference resolution ([Lee et al., 2025](#)). Finally, transmitting sensitive patient data to external models raises privacy concerns, while hosting sufficiently powerful models locally is often impractical.

To address these limitations, we turn to the standard itself. While FHIR defines the data structure, FHIRPath ([HL7 International, 2020](#)) is its native path-based query language. Unlike SQL—which requires backend database access rarely available to patient-facing applications—FHIRPath operates at the API layer and is broadly supported across interoperable health systems.

Adopting a Text-to-FHIRPath approach offers three distinct advantages over RAG or SQL-based methods:

1. **Determinism & Safety:** A generated query acts as an intermediate logic step that can be validated before execution, substantially reducing hallucination at the answer level by shifting risk to symbolic query generation.
2. **Data Minimization:** By executing the query logic locally or at the API level, the system retrieves only the specific data point requested

(e.g., a single observation value) rather than retrieving and processing entire clinical documents. This drastically reduces the context window cost and minimizes personally identifiable information exposure.

3. **Interoperability:** A model trained to generate FHIRPath works across any system compliant with the US Core Data for Interoperability (USCDI) standard, whereas SQL generation is brittle and tied to specific database schema.

Despite the growing need for patient-centric QA, existing benchmarks largely evaluate adjacent but insufficient problem settings. Much prior work targets clinician-facing analytics or structured querying for secondary clinical use (Lee et al., 2022), rather than patient access to their own records. Patient-oriented benchmarks such as archEHR-QA emphasize interpreting verbose patient questions, identifying relevant evidence, and generating grounded answers from free-text clinical notes (Soni and Demner-Fushman, 2026). While critical for modeling patient-clinician communication, these tasks operate at the document level and do not evaluate precise, field-level access to structured EHR data.

Benchmarks that evaluate QA over FHIR often rely on synthetic or heavily curated data (Soni et al., 2019; Kothari and Gupta, 2025), substantially reducing the ambiguity, noise, and scale of real patient records. More recent benchmarks assess agentic interaction with live FHIR servers (Lee et al., 2025), evaluating end-task correctness but without exposing or supervising the executable query logic required to reliably traverse FHIR graphs.

This work proposes a deterministic FHIRPath-based approach as a safer, verifiable alternative to standard retrieval methods. By pairing executable query logic with answers, the approach prioritizes precision and safety over broad recall. The method is evaluated using a novel benchmark derived from MIMIC-IV on FHIR Demo and EHR-SQL (Bennett et al., 2024, 2023; Lee et al., 2022, 2024), leveraging real clinical data rather than synthetic proxies. A patient-perspective paraphrasing pipeline is introduced to assess the system’s ability to translate non-expert language into structured FHIRPath queries. An overview of the primary contributions follows:

- **A Question-FHIRPath-Answer dataset.** FHIRPath-QA combines realistic clinical questions, executable FHIRPath queries, and answers from real MIMIC-IV on FHIR patient data. The full dataset as well as generation code is publicly available for reproducibility and to support future extension.
- **Patient-oriented Paraphrasing.** Real patients do not ask for "tachycardia" status; they

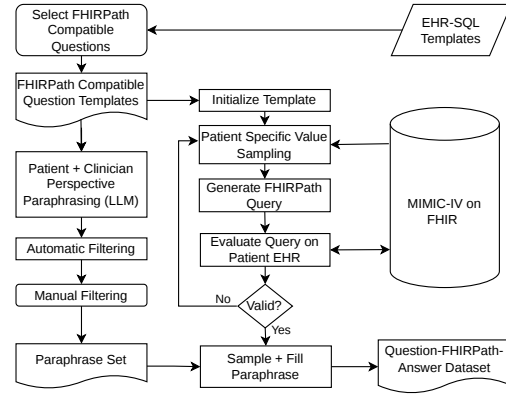


Figure 1: Building FHIRPath-QA involved question selection, LLM-based paraphrasing, and query generation grounded in real patient data.

ask, "Why is my heart racing?" This work includes a pipeline to translate clinical questions into layperson language, evaluating the model’s ability to bridge the semantic gap between vague patient intent and strict FHIR schema.

- **Empirical Evaluation** of retrieval-first and query-first QA pipelines with commercial and fine-tuned models.

2. Related Work

Early EHR QA focused on Text-to-SQL generation for clinician analytics. MIMICSQL (Wang et al., 2020), EHR-SQL (Lee et al., 2022), and EHR-SQL-2024 (Lee et al., 2024) established strong baselines using real clinical questions grounded in the MIMIC-III (Johnson et al., 2016), eICU (Pollard et al., 2018) and MIMIC-IV databases (Johnson et al., 2023). However, these systems rely on direct database access (SQL), which is typically inaccessible from patient-facing applications. Further, these databases have non-standard schema, so transferability to clinical use is limited.

There are limited benchmarks for QA models on FHIR. Soni et al. (2019) curated one of the first question-query-answer data sets in FHIR data. However, they did so using Synthea (Walonoski et al., 2017), a synthetic data-generation pipeline, so the EHRs lack the noise, variability, and scale of real clinical data. Kothari and Gupta (2025) similarly used Synthea to generate a QA dataset for fine-tuning LLMs, but these questions were generated after selecting a small set of relevant resources, further simplifying the task. Consequently, systems are not required to reason over large, longitudinal EHRs with the complexity or scale of real patient records.

More recently, a paradigm shift toward agentic interaction has led to benchmarks that evaluate an LLM’s ability to interact with live FHIR servers. MedAgentBench (Jiang et al., 2025) assesses end-to-end task completion across 10 clinical workflows using 300 physician-authored questions, emphasizing practical correctness. FHIR-AgentBench (Lee et al., 2025) similarly evaluates agentic performance, but focuses specifically on clinician-oriented question answering. Correctness is determined using annotated keys of relevant resources. While this benchmark meaningfully advances QA evaluation on real FHIR data, it does not expose or supervise the executable query logic used to retrieve answers, as the underlying dataset generation relies on SQL. Consequently, the intermediate reasoning required to traverse FHIR graphs is not directly observable or learnable, highlighting the absence of benchmarks that expose executable query logic over real FHIR data.

Taken together, prior work either targets clinician-facing analytics via Text-to-SQL, evaluates FHIR-based QA on synthetic or heavily curated data, or measures agentic interaction with FHIR servers without exposing the underlying executable logic. As a result, none directly supervise the translation of natural language questions into verifiable, executable queries over real-world FHIR data. FHIRPath-QA addresses this gap in Text-to-FHIRPath generation by introducing a benchmark that pairs patient and clinician language questions with executable FHIRPath queries over real MIMIC-IV on FHIR records.

3. Dataset Construction

3.1. Data Source

To ensure questions, queries, and answers are realistic, patient data was used from the MIMIC-IV on FHIR Demo Dataset (Bennett et al., 2024, 2023). Though the released dataset draws from the 100-patient EHRs in MIMIC-IV on FHIR Demo, the generation code extends to the full MIMIC-IV dataset. This data was uploaded into a local instance of the HAPI FHIR server as it is a robust open-source implementation of the HL7 FHIR standard (Hussain et al., 2018).

3.2. Question Templates

Standardized question templates were adapted from EHR-SQL (Lee et al., 2022) to ensure they have clinical relevancy. Templates were selected that focus on individual patients (e.g. "What was the last prescription for patient X?") rather than population-level or aggregate statistics (e.g. "How many patients were diagnosed with hepatitis this year?"). This choice was made because individual

FHIR Resource	Clinical Topics
Patient	Demographics, birth date, gender
Encounter	Hospital visits, ICU stays, admission types, discharge times
Observation	Lab tests, vital signs, weight, intake events, output events, microbiology results
Condition	Diagnoses, clinical status
Medication (Med.)	Drug names
Med.Request	Prescriptions, dosages, drug delivery routes
Med.Administration	Medication administered to patient in hospital
Procedure	Clinical interventions, surgical procedures, timing of events

Table 1: The FHIR resource types used in FHIRPath-QA cover a range of clinical topics required for inpatient care.

patients would not have access to, nor an interest in, other patients’ EHRs. Finally, questions were filtered to ensure they can be answered from the information in FHIR data by a single FHIRPath execution. In total, 61 unique question templates were selected covering a broad range of clinical topics, summarized in Table 1.

3.3. Multi-Perspective Paraphrasing

A multi-perspective paraphrasing strategy was developed to measure the robustness of models across diverse user groups. While existing datasets like EHR-SQL primarily focus on clinician-oriented language, FHIRPath-QA accounts for the varying degrees of medical literacy among potential users of a FHIR-based QA system. The generated paraphrases are divided into two distinct perspectives:

- 1. The Patient/Caregiver Perspective:** Recognizing that patients and caregivers often access their own data through portals, these paraphrases were generated with more conversational and intuitive phrasing. Further, specification of the patient’s ID is removed from the question, since it is assumed that this information would be known from user identification.
- 2. The Clinician Perspective:** This category maintains the technical precision of the original templates. Formal medical terminology is maintained to reflect the language that would be used by a physician or nurse.

Following the observation that modern LLMs can outperform traditional language-generation meth-

```

Bundle.entry.resource.where(
  resourceType='Observation'
  and
  specimen.resolve().type.coding.where(
    (code = 'SEROLOGY/BLOOD')
    or
    (display = 'SEROLOGY/BLOOD')
  ).exists()
  and
  effectiveDateTime.toDateTime() >= @2180-01-01T22:48:00.000Z
  and
  effectiveDateTime.toDateTime() <= @2180-12-31T22:48:00.000Z.highBoundary()
)
.sort(effectiveDateTime).last()
.where(
  hasMember.exists()
  and
  hasMember.resolve().code.coding.system=
    'http://mimic.mit.edu/fhir/mimic/CodeSystem/mimic-microbiology-organism'
).exists()

```

Figure 2: The query generated to answer the question *Were any organisms found in my last SEROLOGY/BLOOD microbiology test this year?* Colored regions indicate logical stages of the query: resource type selection (blue), specimen filtering (green), temporal constraints (orange), and selection of the most recent observation (purple). A final post-aggregation constraint checks that an organism was found (grey).

ods, OpenAI gpt-4o was used for paraphrase generation. For each question template, 50 paraphrases were generated from both the clinician and patient perspectives. The refinement process followed a three-step pipeline:

- 1) All required variables and fillable placeholders in the query templates were verified to be preserved correctly during paraphrasing.
- 2) Lexical diversity was enforced by discarding paraphrases with a pairwise Levenshtein distance less than 10 or a normalized distance under 0.15.
- 3) Semantic alignment was maintained by encoding sentences with the all-MiniLM-L6-v2 transformer and computing cosine similarity. A lower similarity threshold was applied for patient-perspective paraphrases to accommodate the natural linguistic variance of non-physician language.

This process removed over half of the total paraphrases, resulting in a final dataset of 1,239 clinician and 856 patient entries. Patient paraphrases exhibited higher lexical variety and lower semantic similarity compared to clinician paraphrases. For details, see Appendix B.2.

3.4. FHIRPath Query Generation

Consistent with established datasets, the query template and placeholder structure introduced by Lee et al. (2022) are adopted. This framework is extended by mapping question templates to the hierarchical FHIR R4 standard, replacing relational SQL joins with path-based FHIRPath traversals.

FHIRPath expressions were crafted to be executed on a patient bundle containing all resources needed to answer the question. This assumption is valid queries are executed on the server side. For local execution, such a scenario can be replicated

by running queries on a patient’s full bundle: the result of an ‘\$everything’ export.

The questions can be broadly grouped based on four primary response types: counts (e.g. frequency of interventions), existence (true/false), lists (returning multiple entities), and exact responses (retrieving specific names, dosages, or timestamps). Each question type necessitates a distinct set of navigation steps and filter applications to account for the decentralized nature of the FHIR R4 standard, and the different sets of resources relevant to each question. For instance, querying a microbiology test involves navigating the relationship between the event of specimen collection and the event of the result report. These are recorded as distinct clinical observations and require complex link traversal. An example of this is shown in Figure 2.

3.5. Query Validation

The generation pipeline was extensively tested and validated on patient bundles of varying sizes. These tests focused on ensuring syntactic validity of queries by parsing them with an open-sourced, rust-native FHIRPath implementation from OctoFHIR (OctoFHIR). Queries were manually validated for semantic consistency across multiple placeholder configurations.

Further validation focused on execution accuracy over manually annotated FHIR bundles. Since MIMIC-IV on FHIR patient bundles are very large, five smaller versions (averaging 17 resources per bundle) were curated for this task to ensure each question could be answered. Each query type was validated over the five curated patient bundles.

3.6. Dataset Assembly

To ensure questions have answers (e.g. not asking about ICU visits if the patient has never been in the ICU), questions and queries were randomly generated based on the context of specific patients’ entire EHR. To account for date shifts in the de-identification process of MIMIC-IV, datetime filtering was computed with the current date being the date of the last record in the EHR. Queries were then executed against patient bundles to generate the corresponding answers.

3.7. Evaluation Splits

To evaluate different forms of generalization, three complementary evaluation settings were defined.

The dataset is partitioned into training (78%), validation (9%), and test (13%) splits using a stratified split over natural language paraphrases. Paraphrases are disjoint across splits, while question templates and patients may overlap. This setting

Split	Benchmark		Large	
	Clinician	Patient	Clinician	Patient
Train	954 (45.5%)	675 (32.2%)	4721 (38.7%)	4775 (39.1%)
Val	104 (5.0%)	89 (4.2%)	535 (4.4%)	661 (5.4%)
Test	181 (8.6%)	92 (4.4%)	791 (6.5%)	717 (5.9%)
Total	1239 (59.1%)	856 (40.9%)	6047 (49.6%)	6153 (50.4%)
Combined	2095		12200	

Table 2: Dataset Statistics. There are 100 distinct patient records and 61 question templates. The Benchmark contains one question, query, and executed answer per unique paraphrase. FHIRPath-QA Large contains multiple questions and queries per paraphrase, without executed answers. Samples are stratified by split and further categorized by perspective (Clinician vs. Patient). The bottom row displays the combined counts across both perspectives.

evaluates linguistic generalization and is used for all baseline comparisons.

A *query* holdout is defined, in which four entire question templates are excluded from training. These templates correspond to executable FHIRPath queries that do not appear elsewhere in the dataset, while still operating over FHIR resource types observed during training. It accounts for 155 samples in FHIRPath-QA Benchmark. Performance on this split evaluates a model’s ability to synthesize novel executable query logic from natural language, rather than relying on memorized template-level mappings.

To assess structural generalization beyond the observed schema, a *resource* holdout is defined consisting of five templates that require FHIR resources not present in the training data, specifically `Location` and `MedicationAdministration`. This holdout is composed of 126 samples in FHIRPath-QA Benchmark. Success on this split indicates whether fine-tuning enables generalization of FHIRPath syntax and traversal patterns to unseen resource types, despite no exposure to their schemas during training.

3.8. Dataset Makeup

Resources: The FHIRPath-QA dataset is released as a two-tier resource designed to support both model training and reliable execution-based evaluation:

- **FHIRPath-QA Benchmark:** A high-precision evaluation set containing 2,095 question-query-answer triples. Every entry in this subset has been validated via execution against the patient’s ‘\$everything’ bundle to ensure correctness. This serves as an execution-validated benchmark for measuring execution accuracy and linguistic generalization.
- **FHIRPath-QA Large:** A large-scale corpus containing 12,200 question-query pairs. The queries in this dataset are grounded in real patient records, but answers are not provided via

execution. This resource is designed specifically for supervised fine-tuning (SFT), providing the high volume of examples needed for models to learn the complex syntactic nesting of FHIRPath.

Volume and Diversity: In total, the dataset comprises 14,295 unique samples derived from 61 question templates and 2,095 paraphrases (~34.3 per template) which are grounded in 100 real patient EHRs. Each sample is annotated with the paraphrase split label and an indicator for belonging in the unseen query and resource-level holdout groups for simple filtering that ensures large-scale training sets are disjoint from the evaluation benchmarks.

4. Experiments

Experiments were designed and carried out with the following goals in mind: (1) to compare the performance and token-efficiency trade-offs between query-first (text-to-FHIRPath) and retrieval-first approaches; (2) to assess the robustness of these approaches to the variations in patient and clinician-speech; and (3) to investigate the extent to which SFT enables generalization across novel paraphrases, unseen FHIRPath logic, and novel FHIR resource types. All evaluation results in this paper are reported on FHIRPath-QA Benchmark.

4.1. Architectures

Two distinct architectural paradigms for clinical question answering were evaluated.

1. **Ingestion-based Retrieval** approach (referred to as "retrieval") modeled after the single-turn agent in FHIRAgentBench (Lee et al., 2025), requires the LLM to identify and retrieve relevant FHIR resources for a given patient and perform reasoning over the raw content.

2. **FHIRPath query-based execution** approach (referred to as "query-first") tasks the agent with generating a FHIRPath query. This query is then executed against the patient’s complete FHIR bundle to extract an answer, shifting the computational burden from LLM inference over long contexts to deterministic query execution.

4.2. Metrics

Final answer accuracy was used to evaluate model performance. This is assessed as an exact match of the executed result for the query-first approach. Since the retrieval-based approach produces free-text answers, responses were manually evaluated for correctness against the execution-validated benchmark by two independent annotators (the first and second author), both with graduate-level training and relevant domain experience. Any disagreements were resolved through discussion and consensus.

Because all experiments were conducted in the single-turn setting, failures occurred for both approaches. For large patient EHRs, the retrieval-first approach occasionally exceeded the model context limit. For the query-first approach, failures corresponded to invalid FHIRPath syntax. The accuracy excluding these failures was also measured. When failures were included, they were treated as an incorrect answer.

Efficiency was measured via token usage including prompt and completion tokens. For failed retrieval-based attempts, where the token limit was exceeded, the attempted input tokens was recorded.

4.3. Standard Model Comparison

Baseline performance and efficiency were first evaluated across four foundational models: OpenAI o4-mini, 4.1-mini, 4o-mini, and 4.1-nano. The o4-mini model was included due to strong reported performance in similar evaluations (Lee et al., 2025), 4.1-mini for its extended context length, 4o-mini as a broadly capable baseline for comparison, and 4.1-nano for its smaller size. Each model was tested using both retrieval-first and query-first architectures. Evaluation was conducted on the test split of FHIRPath-QA Benchmark, as detailed in Table 2, establishing performance benchmarks for standard large language models (LLMs) on clinical FHIR tasks.

4.4. Fine-tuning Protocol

To evaluate the impact of specialized training, 4o-mini, 4.1-mini, and 4.1-nano models were fine-tuned on the training samples of FHIRPath-QA Large excluding all samples associated with the

unseen query structure and resource-level hold-out templates. Training was conducted through the OpenAI API (1 epoch, batch size 5, learning rate 1.8 for 4o-mini, 2.0 for 4.1-mini, and 0.1 for 4.1-nano). Evaluation was conducted on paraphrases in the test split to assess linguistic generalization (Test 1), on the unseen queries to assess an improved understanding of FHIR elements (Test 2), and finally on unseen resource types to assess resource-level generalization (Test 3).

To explore the effect of the number of patients represented by the training examples, SFT was performed with three subsets of the training dataset containing: 1) 10 patients, 2) 30 patients, and 3) the full 100 patients in MIMIC-IV on FHIR Demo.

5. Results

5.1. FHIRPath-QA Stresses Scalability

Table 3 shows the token usage per question for reasoning and non-reasoning LLMs for the retrieval-first and query-first models. Using o4-mini, the query-first approach yielded a 391x reduction in average tokens per question compared to the retrieval-based approach. The retrieval strategy also exhibited extreme variance in token usage ($SD \approx 1.9M$ tokens), making its running-costs less predictable for clinical deployment. This variance scaled directly with EHR size. In contrast, query-first token usage was stable ($SD \approx 420$ tokens), as the LLM only processes the natural language question and a system prompt. Non-reasoning models had lower token usage, but this had a small impact in the retrieval-first setting, where input tokens dominated.

Strategy	Model	Tokens	SD
Retrieval	o4-mini	629,839	1,926,723
Retrieval	4.1-mini	577,281	1,889,508
Query-first	o4-mini	1,609	420
Query-first	4.1-mini	697	24

Table 3: The retrieval-first approach consumed a substantially higher number of tokens per question than the query-first approach.

5.2. A Challenging Benchmark

Table 4 lists the results of the retrieval-first and query-first approaches with standard LLMs. Overall, the results show that the FHIRPath-QA benchmark poses a challenge to standard LLMs and agentic architectures, with none achieving an overall accuracy above 42%. Across models, o4-mini outperformed 4.1-mini and 4o-mini following the query-first and retrieval-first approaches. When relevant EHRs successfully fit within the context win-

Model	Accuracy		Failure Rate		Acc. Excl. Failures	
	Clinical	Patient	Clinical	Patient	Clinical	Patient
<i>a) Retrieval</i>						
o4-mini	0.33	0.33	0.36	0.34	0.51	0.49
4.1-mini	0.42	0.41	0.17	0.14	0.50	0.48
4o-mini	0.24	0.15	0.46	0.40	0.45	0.25
<i>b) Query-first</i>						
o4-mini	0.35	0.38	0.09	0.15	0.39	0.45
4.1-mini	0.25	0.27	0.15	0.23	0.29	0.35
4o-mini	0.25	0.28	0.02	0.09	0.26	0.31

Table 4: Overall accuracy slightly favors the retrieval-first approach, but this metric is influenced by its higher failure rate. When excluding failures, the retrieval-first approach consistently outperforms the query-first approach. Retrieval-first performs better on clinically phrased questions, whereas the query-first approach performs better on patient-style paraphrases.

dow (i.e. excluding failures), the retrieval-first strategy demonstrated superior baseline performance, which suggests that foundational LLMs lack a thorough understanding of FHIR structures and as a result have a poor ability to convert questions to correct FHIRPath queries.

Notable differences in performance based on question phrasing and system architecture were observed. Retrieval-first scored more highly on clinician over patient paraphrases, which suggests that foundation models rely heavily on lexical overlap. Technical terms in clinician queries may act as anchors to the formal strings found within the FHIR resource JSONs, allowing for more accurate "needle-in-a-haystack" retrieval. Conversely, in the query-first setting, patient-perspective questions yielded higher accuracy.

5.3. Impact of Supervised Fine-Tuning

Results from SFT on the FHIRPath-QA Large corpus indicate that the apparent performance ceiling of the query-first approach stems from base model pretraining rather than architectural limitations. As shown in Figure 3, SFT leads to a dramatic improvement in the model's ability to map natural language to executable logic. For Test 1, which evaluated novel paraphrases of questions seen in training, all models experienced strong gains. Notably, 4o-mini accuracy nearly tripled, rising from a baseline 27% to 79%. This result confirms that specialized training substantially improves the LLMs' ability to recognize diverse linguistic patterns and map them to executable FHIRPath logic.

Test 2 demonstrates model capacity for logical generalization, which utilized entirely novel question templates composed of clinical reasoning elements seen in other contexts. Performance gains across all models in this category suggest that SFT enabled the agents to learn compositional mappings between clinical intent and the FHIR schema, rather than simply memorizing ques-

tion-query pairs.

The results of Test 3 highlight a limit of text-to-FHIRPath generalization. As the models were tasked with querying resources not present in the training set, performance decreased following fine-tuning. While the models become highly proficient in the specific resource types and query patterns represented in the training data, they may overfit to that subset and lose some of the broader, general-purpose schema knowledge present in the base models.

The scale of patient-specific data used during training also proved to be a critical factor in model proficiency. By comparing models fine-tuned on subsets of 10, 30, and 100 patient records, a consistent upward trend in execution accuracy as the diversity of patient data increased was observed. This suggests that the current performance has not yet reached a plateau; rather, it indicates that future work extending the fine-tuning dataset to the full MIMIC-IV-FHIR corpus would likely yield further gains in handling the natural variance and "noise" found in real-world longitudinal records.

Finally, a notable shift occurred in the model's sensitivity to paraphrase perspective post SFT. While the base models performed better on patient paraphrases from their structural simplicity, fine-tuned models consistently reversed this trend, achieving higher accuracy on clinician-paraphrases (see Appendix A.1). This implies that the inherent ambiguity of layperson language characterized by lower semantic similarity to the clinical ground truth imposes a natural performance ceiling on single-turn systems.

5.4. Error Analysis

To evaluate the underlying causes of performance degradation in the query generation pipeline, we performed a qualitative error analysis. Examples of each failure mode are provided in Appendix C.

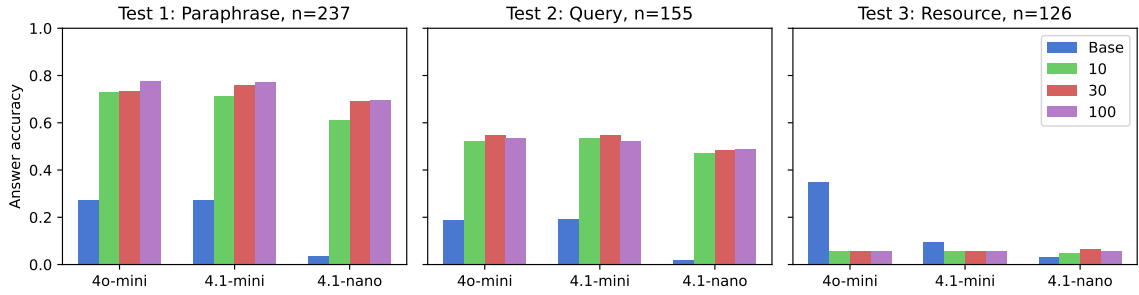


Figure 3: SFT for the query-first approach improved executed answer accuracy on questions with paraphrased inputs over shared query structures (Test 1: Paraphrase) and on questions requiring novel query compositions over FHIR resource types seen during fine-tuning (Test 2: Query). In contrast, performance did not generalize to questions involving unseen resource types (Test 3: Resource). Increasing the number of patients in the fine-tuning dataset (10, 30, 100) led to progressively larger performance gains.

5.4.1. Resource Profile Complexities

A common error mode identified in generated queries stems from an apparent misunderstanding of various FHIR resource profiles. Agents frequently attempt to reference resource elements that do not exist. This failure is indicative that the structure of the resource is not fully understood by the model. For instance, Location resources do not contain temporal information, but some queries tried to access it in search of the time period of a specific visit (Appendix C.2). The correct way to access these types of resources would be to resolve the reference to it, typically via the Encounter resource.

5.4.2. Semantic Misinterpretation

A common failure mode arose when the models were faced with question ambiguity, resulting in syntactically correct FHIRPath that does not correctly address the question. This challenge was more prominent in patient-paraphrased questions lacking precise clinical language. For example, in questions inquiring about a patient’s medications, a particular challenge is in differentiating MedicationRequest resources (prescriptions) and MedicationAdministration (in-hospital drug administration) resources (see Appendix C.3). Resolving this ambiguity often requires clinical knowledge of whether certain drugs are typically administered in hospital or prescribed. Other subtle misinterpretations include counting total events instead of unique events (Appendix C.4) or considering in-hospital transfers as unique visits.

5.4.3. Date Filter Mismatch

Up to 30% of errors in some fine-tuned models were characterized by the generation of the correct FHIRPath query, with the exception of the date filter. In these instances, the model successfully identi-

fies the required resource type and filtering criteria but incorrectly computes the date relative to the current time. An example is provided in Appendix C.1. Quantification of the frequency of this error for each model is provided in Appendix A3.

6. Conclusion

This work presents FHIRPath-QA, a novel dataset and benchmark combining patient-specific questions, validated FHIRPath queries, and executed answers. It is derived from the MIMIC-IV on FHIR Demo dataset, with clinically relevant questions expressed from both clinician and patient perspectives. To our knowledge, this is the first publicly available benchmark and dataset for Text-to-FHIRPath generation grounded in real EHRs.

Empirical comparison of retrieval-first and query-first QA pipelines shows that query-first approaches offer practical advantages in reliability and efficiency when operating over large EHRs. Executing symbolic queries avoids context window limitations, reduces token consumption by several orders of magnitude, and provides a clear and interpretable failure mode when query generation is unsuccessful. At the same time, empirical results demonstrate that retrieval-based approaches remain competitive in answer accuracy when relevant resources can be ingested, underscoring that query-first execution is not a universal replacement for retrieval-based QA.

These findings show that mapping natural language questions to executable FHIRPath logic remains challenging for pretrained LLMs. However, SFT on question–query pairs substantially improves performance and enables generalization to novel paraphrases and previously unseen question compositions over known FHIR resource types.

Importantly, these results reveal a fundamental tension in patient-facing question answering. While fine-tuning improves executable query synthesis,

patient-perspective questions — characterized by greater lexical diversity and semantic ambiguity — impose a natural performance ceiling. Even when models learn FHIRPath syntax, resolving layperson intent into precise clinical constructs remains difficult. This suggests that patient-facing QA may require mechanisms beyond single-turn query generation, such as clarification strategies or schema-aware grounding, to bridge the semantic gap between informal language and structured EHR representations.

By exposing executable query logic over real FHIR data and highlighting the linguistic challenges inherent in patient language, FHIRPath-QA establishes a foundation for safer, more transparent, and more interoperable patient-facing clinical QA systems.

7. Acknowledgements

This work was supported by Smile Digital Health.

8. Limitations

This dataset is constructed from MIMIC-IV on FHIR Demo and primarily reflects ICU and in-patient care from a single institution, which may limit generalizability to other clinical settings. Although patient-oriented paraphrases are included, questions are synthetically generated from templates and may not fully capture the variability and scope of real patient-authored queries. The scope of FHIRPath-QA is restricted to factual, record-grounded questions that can be answered deterministically from structured FHIR resources; interpretive or explanatory clinical questions are out of scope. Finally, SFT experiments cover a limited set of resource types, and reduced performance on unseen resources suggests potential over-specialization.

9. Bibliographical References

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:e53636.

Alex Bennett, Joshua Wiedekopf, Hannes Ulrich, Philip van Damme, Piotr Szul, John Grimes, and Alistair Johnson. 2024. [MIMIC-IV on FHIR](#). *PhysioNet*. Version 2.1.

Alex M Bennett, Hannes Ulrich, Philip van Damme, Joshua Wiedekopf, and Alistair E W Johnson. 2023. [MIMIC-IV on FHIR: converting a decade](#)

[of in-patient data into an exchangeable, interoperable format](#). *Journal of the American Medical Informatics Association*, 30(4):718–725.

Adrian Bussone, Simone Stumpf, and Dymna O’Sullivan. 2015. [The role of explanations on trust and reliance in clinical decision support systems](#). In *2015 International Conference on Healthcare Informatics*, page 160–169. IEEE.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. [Benchmarking large language models in retrieval-augmented generation](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI’24/IAAI’24/EAAI’24. AAAI Press.

Guilherme Del Fiore, T. Elizabeth Workman, and Paul N. Gorman. 2014. [Clinical questions raised by clinicians at the point of care: A systematic review](#). *JAMA Internal Medicine*, 174(5):710.

David A. Hanauer, Qiaozhu Mei, James Law, Ritu Khanna, and Kai Zheng. 2015. [Supporting information retrieval from electronic health records: A report of university of michigan’s nine-year experience in developing and using the electronic medical record search engine \(EMERSE\)](#). *Journal of Biomedical Informatics*, 55:290–300.

HL7 International. FHIR: Fast healthcare interoperability resources. <https://www.hl7.org/fhir/>. Accessed: 2026-01-22.

HL7 International. 2020. HL7 cross-paradigm specification: FHIRPath, release 1. <https://hl7.org/fhirpath/N1/>. Accessed: 2026-01-22.

Mohannad A Hussain, Steve G Langer, and Marc Kohli. 2018. [Learning hl7 FHIR using the HAPI FHIR server and its use in medical imaging with the SIIM dataset](#). *Journal of Digital Imaging*, 31(3):334–340.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025. [Medagentbench: A virtual ehr environment to benchmark medical llm agents](#). *NEJM AI*, 2(9).

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, 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).

- 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).
- Sara Kothari and Ayush Gupta. 2025. Question answering on patient medical records with private fine-tuned LLMs. *arXiv preprint arXiv:2501.13687*.
- Gyubok Lee, Elea Bach, Eric Yang, Tom J. Pollard, Alistair Johnson, Edward Choi, Yugang Jia, and Jong Ha Lee. 2025. [FHIR-AgentBench: Benchmarking LLM agents for realistic interoperable EHR question answering](#). In *Proceedings of the Machine Learning for Healthcare Conference (ML4H)*.
- Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. 2022. [EHRSQL: A practical text-to-SQL benchmark for electronic health records](#). In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, page 15589–15601. Neural Information Processing Systems Foundation, Inc. (NeurIPS).
- Gyubok Lee, Sunjun Kweon, Seongsu Bae, and Edward Choi. 2024. [Overview of the ehrsql 2024 shared task on reliable text-to-sql modeling on electronic health records](#). In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, page 644–654. Association for Computational Linguistics.
- 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.
- OctoFHIR. [fhirpath-rs: Fhirpath tools written in rust](#). <https://github.com/octofhir/fhirpath-rs>. Accessed: 2025-07-22.
- World Health Organization et al. 2023. WHO and HL7 collaborate to support adoption of open interoperability standards. *World Health Organization. Recuperado el*, 10.
- Tom J. Pollard, Alistair E. W. 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).
- Lisette C. Roman, Jessica S. Ancker, Stephen B. Johnson, and Yalini Senathirajah. 2017. [Navigation in the electronic health record: A review of the safety and usability literature](#). *Journal of Biomedical Informatics*, 67:69–79.
- Sarvesh Soni and Dina Demner-Fushman. 2026. [ArchEHR-QA: A Dataset for Addressing Patient’s Information Needs related to Clinical Course of Hospitalization](#). *PhysioNet*. Version 1.3.
- Sarvesh Soni, Meghana Gudala, Daisy Zhe Wang, and Kirk Roberts. 2019. [Using FHIR to construct a corpus of clinical questions annotated with logical forms and answers](#). In *AMIA Annual Symposium Proceedings*, pages 1207–1215. American Medical Informatics Association.
- Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2017. [Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record](#). *Journal of the American Medical Informatics Association*, 25(3):230–238.
- Ping Wang, Tian Shi, and Chandan K. Reddy. 2020. [Text-to-sql generation for question answering on electronic medical records](#). In *Proceedings of The Web Conference 2020*, WWW ’20, page 350–361. ACM.
- Jiajie Zhang and Muhammad Walji. 2014. [Better EHR: usability, workflow and cognitive support in electronic health records](#).
- Wan Zhang and Jing Zhang. 2025. [Hallucination mitigation for retrieval-augmented large language models: A review](#). *Mathematics*, 13(5):856.

A. Additional Tables

A.1. Comprehensive Performance Results

See [Table A1](#).

A.2. Comprehensive Usage Results

See [Table A2](#).

A.3. FHIRPath Query Match Rates

See [Table A3](#).

B. Additional Figures

B.1. Token Usage by Patient EHR Size

See [Figure B1](#).

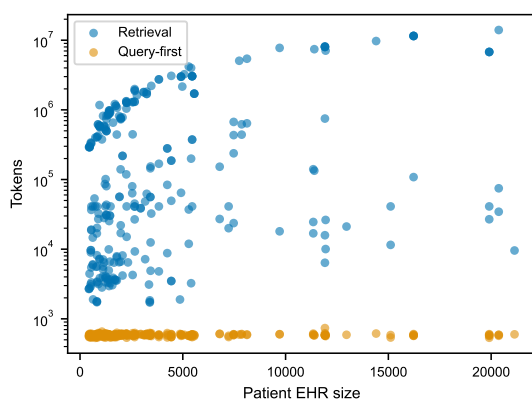


Figure B1: Patient record size vs token usage (log scale) using (o4-mini). The upper bound on tokens used in the retrieval-first approach scales linearly with the patient EHR size, whereas query-first remains constant. Retrieval tokens corresponds to total tokens or attempted tokens if the context length limit was exceeded.

B.2. Paraphrase Statistics

See [Figure B2](#).

C. Error Analysis Examples

C.1. Example 1

Model: query_gen-4.1-nano_100

Question: until last year, for patient 10003400, what was the last time a medication was prescribed using the IV route?

Context: Assume current datetime is 2137-09-02T19:17:11

true_query: Bundle.entry.resource.where(resourceType='MedicationRequest'anddosageInstruction.route.c

oding.code='IV'andauthoredOn.toDateTime()<=@2136-12-31T19:17:11.000Z).sort(authoredOn).last().authoredOn.toDateTime()

model_query: Bundle.entry.resource.where(resourceType='MedicationRequest'anddosageInstruction.route.coding.code='IV'andauthoredOn.toDateTime()<=@2137-12-31T19:17:11.000Z).sort(authoredOn).last().authoredOn.toDateTime()

true_answer: 2136-12-14T08:52:22-05:00

model_answer: 2137-09-02T13:45:41-04:00

Analysis: The model used the incorrect date in the generated query despite having an exact match otherwise.

C.2. Example 2

Model: query_gen-4o-mini_100

Question: Can you tell me what the last careunit was until 2144?

Context: Assume current datetime is 2145-01-03T20:33:00

true_query: Bundle.entry.resource.where(resourceType='Encounter'andselect((period.start|period.end).sort().first().toDateTime())<=@2144-12-31T05:02:42.000Z).sort(period.start).last().location.sort(period.start).last().location.resolve().name

model_query: Bundle.entry.resource.where(resourceType='Location'andname.exists()andperiod.start.toDateTime()<=@2144-12-31T23:25:29.000Z).sort(period.start).last().name

true_answer: Cardiac Vascular Intensive Care Unit (CVICU)

model_answer: []

Analysis: The generated query does not demonstrate knowledge about the structure of a Location or Encounter resource. Location resources contain no temporal information; they are linked to Encounters which contain visit information including the time period.

C.3. Example 3

Model: query_gen-4.1-mini_100

Question: I'd like to know the number of times I took Potassium Chloride until 20/05/2187.

Context: Assume current datetime is 2187-05-23T17:20:00

true_query: Bundle.entry.resource.where(resourceType='MedicationAdministration'andcontext.resolve().identifier.system='http://fhir.mimic.mit.edu/identifier/encounter-icu'andmedicationCodeableConcept.coding.display='PotassiumChloride'andselect((effectiveDateTime|effectivePeriod.start).sort().first().toDateTime())<=@2187-05-20T18:21:11.000Z).count()

model_query: Bundle.entry.resource.where(resourceType='MedicationRequest'and(medicationReference.resolve().identifier.where(value='PotassiumChloride').e

Model	Base/FT	Accuracy		Failure Rate		Acc. Excl. Failures	
		Clinical	Patient	Clinical	Patient	Clinical	Patient
a) Retrieval							
4.1-mini	Base	0.42	0.41	0.17	0.14	0.50	0.48
4o-mini	Base	0.24	0.15	0.46	0.40	0.45	0.25
o4-mini	Base	0.33	0.33	0.36	0.34	0.51	0.49
b) Query-first							
4.1-mini	Base	0.25	0.27	0.15	0.23	0.29	0.35
4.1-mini	FT-10	0.68	0.61	0.02	0.10	0.69	0.67
4.1-mini	FT-30	0.72	0.64	0.01	0.09	0.73	0.70
4.1-mini	FT-100	0.73	0.66	0.01	0.09	0.74	0.73
4o-mini	Base	0.25	0.28	0.02	0.09	0.26	0.31
4o-mini	FT-10	0.70	0.62	0.01	0.09	0.70	0.68
4o-mini	FT-30	0.70	0.63	0.02	0.09	0.71	0.69
4o-mini	FT-100	0.73	0.66	0.01	0.09	0.74	0.73
4.1-nano	Base	0.02	0.07	0.22	0.29	0.02	0.09
4.1-nano	FT-10	0.59	0.49	0.04	0.11	0.61	0.55
4.1-nano	FT-30	0.68	0.54	0.04	0.11	0.71	0.61
4.1-nano	FT-100	0.67	0.57	0.01	0.11	0.68	0.63
o4-mini	Base	0.35	0.38	0.09	0.15	0.39	0.45

Table A1: Comprehensive experimental results of retrieval and query-first strategies across clinical and patient perspectives for all models evaluated on the test split. FT-X indicates the fine-tuning dataset included X patients.

Model	Base/FT	Token Usage	
		Tokens / Question	STD
a) Retrieval			
4.1-mini	Base	577,281	1,889,508
4o-mini	Base	737,846	1,992,794
o4-mini	Base	629,839	1,926,723
b) Query-first			
4.1-mini	Base	697	24
4.1-mini	FT-10	704	25
4.1-mini	FT-30	704	25
4.1-mini	FT-100	703	25
4o-mini	Base	692	18
4o-mini	FT-10	703	25
4o-mini	FT-30	704	25
4o-mini	FT-100	703	25
4.1-nano	Base	684	19
4.1-nano	FT-10	703	25
4.1-nano	FT-30	703	25
4.1-nano	FT-100	702	23
o4-mini	Base	1,609	420

Table A2: Token consumption per question on the test split across retrieval and query-first strategies for all models evaluated. FT-X indicates the fine-tuning dataset included X patients.

xists()ormedicationCodeableConcept.coding.code='PotassiumChloride')andauthoredOn.toDateTime()<=@2187-05-20T19:10:10.000Z).count()

true_answer: 1

model_answer: 0

Analysis: The patient-paraphrased question contains ambiguity in the phrase "I took Potassium Chloride". The incorrectly assumes Potassium Chloride was a prescription, but it was intended to

be a question about drug intake while an in-patient. The clinician-paraphrased question would clearly specify if the drug was 'prescribed' or 'administered'.

Model (Train Size)	Query Match Rate	Query Match Rate with Masked Numeric Quantities
4.1-mini (Base)	0.01	0.01
4.1-mini (10)	0.31	0.58
4.1-mini (30)	0.35	0.63
4.1-mini (100)	0.34	0.63
4.1-nano (Base)	0.00	0.01
4.1-nano (10)	0.07	0.39
4.1-nano (30)	0.14	0.44
4.1-nano (100)	0.16	0.43
4o-mini (Base)	0.03	0.05
4o-mini (10)	0.24	0.58
4o-mini (30)	0.33	0.63
4o-mini (100)	0.35	0.64
o4-mini (Base)	0.03	0.03

Table A3: The query match rate with masked numeric quantities consistently increased after fine-tuning, demonstrating that a high number of query inaccuracies were due to incorrect date filtering in otherwise correct queries. This metric was computed by masking all numeric substrings (i.e. dates, patient ids, dose quantities) in the generated and ground truth query before comparing strings.

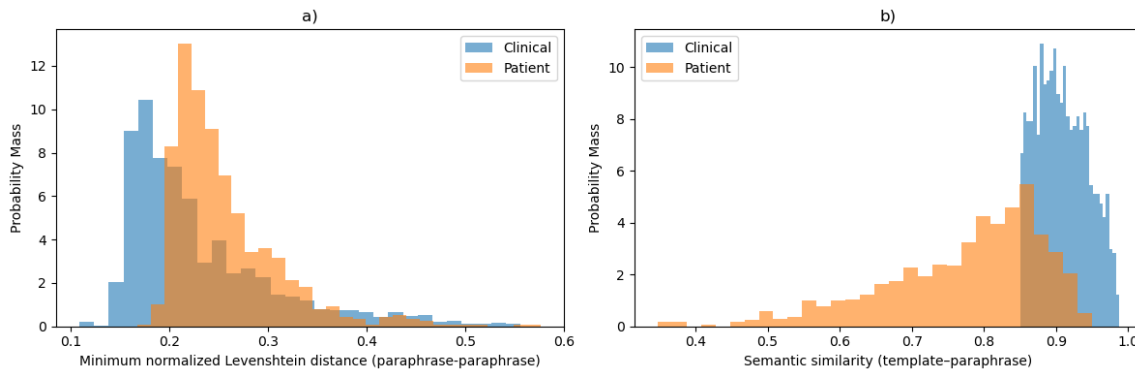


Figure B2: Histograms of a) the minimum normalized Levenshtein distance between paraphrase and each other paraphrase of the same question template and b) the semantic similarity of paraphrase to original question template by perspective.

C.4. Example 4

Model: query_gen-4.1-mini_100

Question: I need to know how many medications were issued to me until 4 months ago.

Context: Assume current datetime is 2151-01-15T23:17:00

true_query: Bundle.entry.resource.where(resourceType='MedicationRequest'andauthoredOn.toDateTime()<=@2150-09-30T23:17:00.000Z).select(medicationReference.resolve())|medicationCodeableConcept).distinct().count()

model_query: Bundle.entry.resource.where(resourceType='MedicationRequest'andauthoredOn.toDateTime()<=@2150-09-30T23:17:00.000Z).count()

true_answer: 258

model_answer: 1442

Analysis: The generated query counts the total number of prescriptions instead of counting distinct medication types as the question required. The result is a partially correct query.

C.5. Example 5

Model: query_gen-4o-mini_base

Question: Which microbiology test did patient 10003046 undergo last since 14 months ago?

Context: Assume current datetime is 2154-01-09T11:53:00

true_query: Bundle.entry.resource.where(resourceType='Observation'andcode.coding.system='http://fhir.mimic.mit.edu/CodeSystem/microbiology-test'andeffectiveDateTime.toDateTime()>=@2152-11-01T11:53:00.000Z).sort(effectiveDateTime).last().code.coding.display

model_query: Bundle.entry.resource.where(resourceType='Observation'andcode.coding.display='Microbiology'andeffectiveDateTime.toDateTime()>=@2152-11-01T11:53:00.000Z).sort(effectiveDateTime).last().code.coding.display

true_answer: C. difficile PCR

model_answer: ""

Analysis: The generated query fails to specify a microbiology test code system. It incorrectly applies

a generic display filter that does not exist, missing the required microbiology result. Apart from this filter clause, the query is correct.

C.6. Example 6

Model: query_gen-4.1-mini_base

Question: Provide the number of Condom Cath outputs associated with patient 10035631 since 10 years ago

Context: Assume current datetime is 2116-03-12T07:45:00

true_query: Bundle.entry.resource.where(resourceType='Observation'andcode.coding.display='CondomCath'andeffectiveDateTime.toDateTime())>=@2106-01-01T07:45:00.000Z).count()

model_query: Bundle.entry.resource.where(resourceType='Observation'andcode.coding.display='CondomCath'andeffectiveDateTime.toDateTime())>=@2106-03-12T07:45:00.000Z).count()

true_answer: 1

model_answer: 1

Analysis: The generated query uses the incorrect date component, making the correct evaluated answer coincidental.

C.7. Example 7

Model: query_gen-o4-mini_base

Question: How many times was patient 10000032 given Sodium Polystyrene Sulfonate until 2180?

Context: Assume current datetime is 2180-08-10T22:48:00

true_query: Bundle.entry.resource.where(resourceType='MedicationRequest'and(medicationReference.resolve().identifier.where(value='SodiumPolystyreneSulfonate').exists()ormedicationCodeableConcept.coding.code='SodiumPolystyreneSulfonate')andauthoredOn.toDateTime())<=@2180-12-31T22:18:53.000Z).count()

model_query: Bundle.entry.resource.where(resourceType='MedicationAdministration'andmedicationCodeableConcept.coding.display='SodiumPolystyreneSulfonate'andeffectiveDateTime.toDateTime())<=@2180-12-31T22:48:00.000Z).count()

true_answer: 4

model_answer: 0

Analysis: The generated query has no correct components. Likely stemming from ambiguity in the question with the phrase "given Sodium Polystyrene Sulfonate", the model assumes this to be a drug administered in a hospital context.

COACH Meets QUORUM: A Framework and Pipeline for Aligning User, Expert, and Developer Perspectives in LLM-Generated Health Counselling

Yee Man Ng*, Bram van Dijk†, Pieter Beynen◇, Otto Boekesteijn◇,
Joris Jansen◇, Gerard van Oortmerssen*, Max van Duijn*, Marco Spruit*†

*Leiden University, {y.m.ng, m.r.spruit, g.van.oortmerssen, m.j.van.duijn}@liacs.leidenuniv.nl

†Leiden University Medical Center, b.m.a.van_dijk@lumc.nl

◇Healthy Chronos, {pieter.beynen, otto.boekesteijn, joris.jansen}@healthychronos.com

Abstract

Systems that collect data on sleep, mood, and activities can provide valuable lifestyle counselling to populations affected by chronic disease and its consequences. Such systems are, however, challenging to develop; in addition to *reliably* extracting patterns from user-specific data, systems should contextualise these patterns with validated medical knowledge to ensure the *quality* of counselling and generate counselling that is *relevant* to a real user. We present QUORUM, an evaluation framework that unifies these developer-, expert-, and user-centric perspectives, and show with a real case study that it meaningfully tracks convergence and divergence in stakeholder perspectives. We also present COACH, a Large Language Model-driven pipeline to generate personalised lifestyle counselling for our *Healthy Chronos* use case, a diary app for cancer patients and survivors. Applying our framework indicates that, overall, users, medical experts, and developers converge on the view that the generated counselling is relevant, of good quality, and reliable. However, stakeholders also diverge on the tone of the counselling, sensitivity to errors in pattern-extraction, and potential hallucinations. These findings highlight the importance of multi-stakeholder evaluation for consumer health language technologies and illustrate how a unified evaluation framework can support trustworthy, patient-centered NLP systems in real-world settings.

Keywords: consumer health question answering, large language models in health, remote monitoring

1. Introduction

People living with the consequences of cancer (treatment) often suffer from long-term effects such as fatigue and mental health issues (Akechi et al., 1999). While it is known that healthy lifestyle changes like increasing physical activity positively impact health outcomes and quality of life, for both cancer patients and survivors (Hoedjes et al., 2022), they currently experience the state of lifestyle counselling in healthcare as not readily available, too fragmented, and general (van Aken et al., 2025; Drbohlav Ollerton et al., 2025; Tuinman et al., 2024). This demonstrates the need for personalised lifestyle counselling, through reliable pattern extraction from user-specific health records and with validated sources, in order to produce high-quality and evidence-based answers to consumer-health questions. While many AI-driven systems that generate counselling exist, their evaluation methodologies are often fragmented and lack realistic contexts (Lai et al., 2025; Raji et al., 2021).

This is why this paper contributes **QUORUM** (**Q**uality, **O**utcome **R**eliability, and **U**ser-relevance from **M**ultiple stakeholders), a new evaluation framework that aligns with the goals and interests of three key types of stakeholders in the development of NLP systems for health. From a *user-centric* perspective, counselling is *relevant* to users if the

advice aligns with their situation, increases the likelihood of action based on the counselling, and if they are satisfied with the tone and length of the counselling. *Experts*, on the other hand (in our case, cancer information specialists from kanker.nl, henceforth ‘experts’) focus on the *quality* of the counselling with respect to its correctness, and the appropriate tone and length for the population and disease they know well. *Developers/researchers* (two authors, henceforth ‘developers’) focus on a more technical evaluation and have as a vantage point *reliability*, i.e., baseline performance on key metrics such as faithfulness and completeness of the counselling, and hallucination rates. We argue that these three perspectives must be included in the evaluation of AI systems that support specific medical populations. Since we cannot expect such stakeholders to evaluate the output of AI systems by the same standards, they provide irreducible, complementary perspectives that only jointly indicate whether a given system is likely to have a positive, real-world impact.

We show QUORUM in action with a concrete use case: the Dutch *Healthy Chronos* application. This is a digital diary app that aids cancer patients and survivors by tracking sleep, activities, mood, and custom user-defined goals over time. We develop the Large Language Model (LLM) driven pipeline **COACH** (**C**ontextualised **O**utcome-

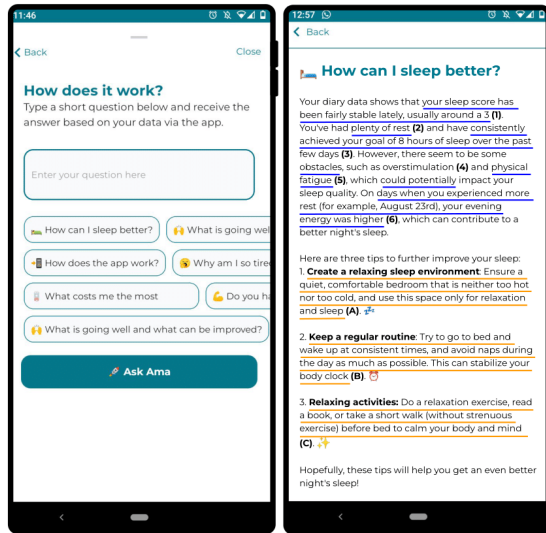


Figure 1: Left: user interface for submitting queries. Right: generated counselling for query ‘How can I sleep better?’ Blue underlining illustrates *claims about the user data*; orange indicates *contextualisation statements* that contextualise found patterns with information from the knowledge database. Claims and statements are labelled with (1), (A), etc., but this is hidden from the user.

Adaptive Counselling for Health) for *Healthy Chronos*, which uses Retrieval Augmented Generation (RAG) (Lewis et al., 2020) to generate personalised lifestyle counselling. The pipeline ingests a user query about their data (e.g. *How can I sleep better?*), after which COACH retrieves relevant chunks for that query from the external, validated knowledge database *kanker.nl*. COACH then contextualises the patterns extrapolated from the user data with the retrieved chunks into personalised counselling (for an impression, see Figure 1). By using QUORUM on COACH, we demonstrate that relevant, good-quality, and reliable lifestyle counselling is possible with LLMs. Nineteen Dutch app users acknowledge that the counselling aligns with their situation, they intend to follow up on the counselling, and endorse its tone and length. Six independent experts from *kanker.nl* deem the counselling of good quality in terms of the correctness of the contextualisation of extracted user patterns, and its tone and length. Technical evaluation by two developers shows COACH is reliable, as most claims (79%) are consistent with user data; moreover, for virtually all claims (97%), the relevant information has been retrieved. Still, we also find that 22% of the generated counselling statements are not strictly traceable to the knowledge database, thus are hallucinations (though not of a harmful kind).

Besides this convergence, QUORUM also highlights divergence between stakeholders. For exam-

ple, users and experts differ on the desired tone of the response, and differ with developers in the sensitivity to errors and hallucination in the reported patterns and counselling. These results highlight the utility of QUORUM and COACH for other populations and medical conditions, by helping build systems with a larger real-world impact.

2. Background

Consulting the internet for information on lifestyle change to improve one’s health and quality of life is increasingly common (Raeside et al., 2022). Still, people struggle with finding information fitting their specific queries, due to answer complexity, potentially incorrect information, and the lack of alignment with their situation (Bondarenko et al., 2021; Pugachev et al., 2023; Kayser et al., 2015). Though LLMs are recognised as potentially useful for providing lifestyle counselling due to their language and reasoning abilities, a recent overview shows that particularly evaluation methodologies for LLM-driven lifestyle counselling are fragmented, weak, and lack realistic contexts (Lai et al., 2025).

Related work has further shown that for LLM-driven chatbots in health, stakeholders can have diverging perspectives relevant to system development. For example, whereas users may prioritize the tone of the interaction, professional caregivers may emphasize safety and clinical validation aspects (Van Dijk et al., 2025). Still, current work typically targets only a subset of stakeholders, like relevance from a user perspective, but no further quality assessment of system output (Meywirth, 2024), or expert evaluation without including real users (Merrill et al., 2024). While such studies address important issues, often with novel AI techniques, extrapolating findings to real populations remains hard.

However, a study close to our setup is Steenstra et al. (2024), and it concerns an LLM-driven tool for Motivational Interviewing regarding alcohol use. It also adopts a three-way evaluation with different stakeholders, where humans evaluated the tone of the LLM-generated outputs, researchers evaluated the presence of appropriate MI techniques in the responses, and independent experts overall tested the system by role-playing interactions. Yet, Steenstra et al.’s work differs from ours in that their tool is not embedded in the real world.

Regarding system development, research has shown that many LLM-driven systems for lifestyle counselling still rely on proprietary models and cloud services (Lai et al., 2025). For startups, this is perhaps understandable, as self-hosting LLMs is costly, while cloud access to capable LLMs is cheap (Nandagopal, 2025); these are key considerations for startups like our *Healthy Chronos*

use case that inform some methodological choices we make below.

3. Methodology

3.1. QUORUM Framework

QUORUM unifies three complementary perspectives from the literature. First, the **model-centric perspective** that emphasises well-known standardized metrics (e.g., recall) on common tasks (e.g., retrieving relevant elements). These metrics are important as they allow quickly probing a system’s base performance and comparison with other systems, but they are poor proxies for a system’s real-world impact (Wang et al., 2024a; Li et al., 2025). This is nevertheless a common perspective to adopt for AI **developers**, and in our framework, it involves faithfulness, completeness, and hallucination metrics as common to RAG evaluation (Papageorgiou et al., 2025). Faithfulness concerns the extent to which generated claims about the user data are consistent with that user’s data; completeness concerns the extent to which all relevant information in the user data for some query was retrieved; hallucination concerns the existence of contextualisation statements in the counselling that cannot be traced to the knowledge database (Recall the difference between *claims about user data* and *contextualisation statements* from Figure 1). Rather than factual incorrectness (the more common definition of hallucination), we consider lifestyle counselling that might be fitting but not grounded in the knowledge database, i.e., any information generated not from the parametric knowledge of the model, to be a type of hallucination.

The **user-centric perspective**, on the other hand, focuses on how well model behaviour actually helps a **real user** to reach a specific goal, thus it is about the relevance of a system’s output (Wang et al., 2024b). We asked users to evaluate the alignment of the response with their personal situation, the tone and length of the response, and the likelihood that the user would follow up on the counselling provided. These more subjective measures aim to capture the relevance of the response to the user.

We also incorporate the **expert-centric perspective**. In a health context, besides user and developer, **medical experts** as stakeholders often have different concerns (Van Dijk et al., 2025), for example, about the overall adequacy of the medical information in the response, because actions may be based on it, and because users may re-use the generated information in real medical consultations. Thus, we ask experts to evaluate the correctness of the contextualisation, based on their expertise on the (consequences of) cancer, as well as their

opinion on the style and length of the response.

3.1.1. Holistic, Iterative Evaluation

In this way, we aim to incorporate a holistic perspective on the generated counselling that strikes a good balance between technical (model-centric), subjective (user-centric), and medical (expert-centric) aspects. Combining these three perspectives is vital, as single perspectives cannot by themselves provide a good estimation of the real-world impact of AI systems in health; they are complementary, and one perspective cannot be reduced to another. Though they do overlap in topics, the perspectives try to capture those opinions that can be reasonably expected from different stakeholders, given their background, goals, and interests.

For example, users could be less likely to verify reported patterns or lifestyle advice if an answer is in line with what they already believe; this is a common issue in dealing with the fluent output of LLMs nowadays and known as confirmation and consistency bias (Nirman et al., 2025). Or, they could be less likely to do so due to automation bias: an over-reliance on the capabilities of algorithms in generating valid information (Boonprakong, 2025). This shows how model- and expert-centric evaluation can complement each other. On the other hand, an AI system with perfect reliability may still have little real-world impact because few of its outputs may be meaningful to a user; here, the user-centric perspective provides the best proxy for real impact. What is still missing, then, is some assurance that the reliable and relevant output is also of good quality from an expert-centric perspective; only experts can adequately assess medical content.

QUORUM was conceived of as a quick probe in the context of *iterative* system development, where, especially for interactive systems, many brief stakeholder evaluations need to be done to incorporate their views in future versions of the system (van Gemert-Pijnen et al., 2011; Van Dijk et al., 2025). After evaluation is done with each group, a QUORUM user can compute, per perspective, a grand average that indicates how well the system at issue is doing from that perspective, while the individual variables provide details on the perspective’s key aspects. See Table 1 for a schematic overview of variables in QUORUM; we further show and discuss outputs of our framework in section 4.

3.1.2. Data Collection

Here we describe the procedures for data collection from users, experts, and researchers. All protocols were approved by the Leiden University Ethics Review Board. Though we share the source code of our pipeline (see subsection 3.2), we cannot share user data itself nor `Healthy Chronos`

Perspective (N)	Statement evaluated	Variable	Measurement	Interpretation
Users (19)	1 <i>The response aligns with my situation</i>	Alignment	5-point Likert scale	Avg. score indicating alignment with user's situation (mood, energy, etc.)
	2 <i>I will follow the tips in the response</i>	Follow-up	5-point Likert scale	Avg. score indicating whether user is likely to act on the response
	3 <i>The response has the right tone</i>	Tone	5-point Likert scale	Avg. score of response tone quality
	4 <i>The response has the right length</i>	Length	5-point Likert scale	Avg. score of response length
Experts (6)	1 <i>The contextualisation is correct</i>	Correctness	5-point Likert scale	Avg. score of contextualisation quality
	2 <i>The response has the right tone</i>	Tone	5-point Likert scale	Avg. score of response tone quality
	3 <i>The response has the right length</i>	Length	5-point Likert scale	Avg. score of response length
Developers (2)	1 <i>This claim is consistent with the data</i>	Faithfulness	Binary	% of claims supported by user data
	2 <i>For this claim, the relevant user variables were retrieved</i>	Completeness	Binary	% of claims for which relevant columns were retrieved from user data
	3 <i>This contextualisation statement can be traced to the relevant chunks found by the pipeline</i>	Hallucination	Binary	% of claims grounded in the knowledge database (<code>kanker.nl</code>)

Table 1: Contents (translated from Dutch) of the QUORUM evaluation framework. Five-point Likert scales are structured as *strongly disagree* (1), *disagree* (2), *neutral* (3), *agree* (4), *strongly agree* (5) .

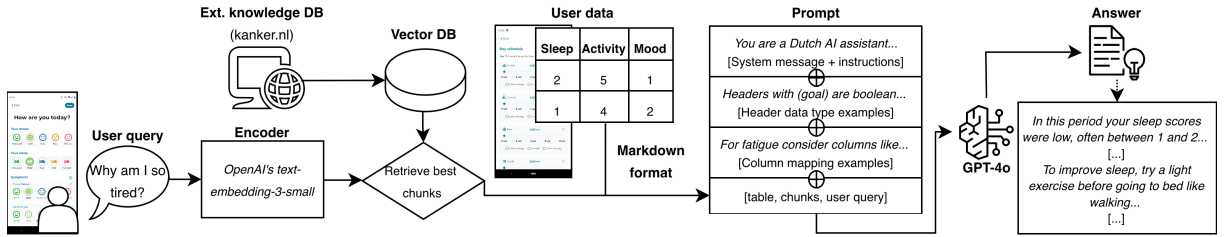


Figure 2: Overview of the pipeline flow and inputs.

source code due to privacy and intellectual property considerations.

Users Nineteen Dutch users were recruited via the app to participate in this study. As shown in Figure 1, users can query their data, but the app also suggests query topics based on frequency. In our evaluation, users evaluate responses on the question: *What is going well and what can be improved?* One reason for this is that this query covers a large part (41%) of a random sample of 100 queries. Also, this query has a broad scope, meaning that the pipeline is forced to correctly retrieve and analyse potentially many different relevant variables, which can be more challenging compared to more specific queries (e.g., about sleep). Users received the query and generated a response in the app, and then completed a short survey on the statements given in Table 1. They were asked to evaluate the full response, and were also invited to submit open remarks for each query-response pair. This amounted to nineteen evaluated query-response pairs. We did not collect further demographic data due to privacy considerations, but the user base of *Healthy Chronos* contains mostly Dutch adolescents and young adults.

Experts Six experts from `kanker.nl` consented to evaluating in total 50 random user query-

response pairs, and they evaluated for each pair the four statements given in Table 1. Experts were instructed to focus on the correctness of the *contextualisation* (see Figure 1, as due to privacy concerns it was not feasible to share user data with them, and because the contextualisation is based on the knowledge database (`kanker.nl`) they were familiar with. Besides providing scores, experts were also invited to submit remarks at every query-response pair.

Developers Two developers evaluated in two rounds in total 66 *claims about the user data* and 31 *contextualisation statements* (see Figure 1) from 10 random query-response pairs. Statements 1 and 2 in Table 1 were used for the claims about user data, whereas statement 3 was used for the contextualisation. After the developers initially evaluated the whole set independently, disagreements were discussed to consensus afterwards.

We computed measures for interrater reliability for experts and developers, for various variables, and therefore describe these metrics in tandem with our results in section 4.

3.2. Pipeline

Healthy Chronos is a digital diary for people living with the consequences of cancer (treatment),

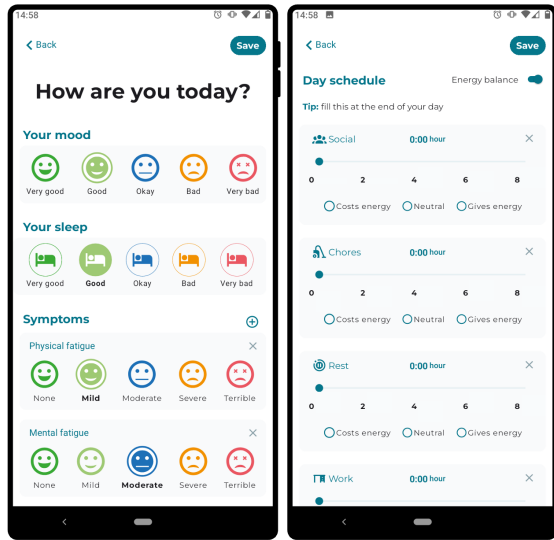


Figure 3: User interface of Healthy Chronos.

developed in cooperation with the Netherlands Cancer Institute and Dutch Cancer Society. The goal is to help this group cope with the physical and mental effects of dealing with cancer (treatment), though the app can accommodate any user-specified goal. Through reminders and surveys, the app helps users to score variables like mood, sleep, and energy (scores on five-point scales), activities like work, social activities, chores, and exercise (in hours), and goals, like resting at least 30 minutes (booleans) (Figure 3). Users receive push messages to submit scores at the end of the day, and users can submit arbitrary queries about their data. Healthy Chronos has a free version and one free year of premium use through a waiver.¹

COACH was developed as an efficient Python pipeline to quickly generate counselling from user queries within the Healthy Chronos app. It was fully developed and evaluated in the Azure environment of Healthy Chronos, in line with the app’s privacy and security agreements with users. COACH’s full code and prompt are available on GitHub.²

COACH’s workflow is as follows (also see Figure 2). Once a single user query (or batch) is submitted to the application, COACH embeds the user query using the *text-embedding-3-small* encoder from OpenAI. COACH then scrapes a subset of Dutch webpages from *kanker.nl* that contain trustworthy information on cancer and lifestyle (advice) written by cancer information experts, on topics like how to deal with fatigue or talking about the disease with loved ones. It then converts them to

a vector store so that knowledge stays up to date. Thereafter, COACH retrieves four chunks most relevant to the query. COACH transforms the table and then appends it with chunks and a query to a prompt. Prompt design followed the following insights from recent work:

- Separating instructions and examples from ‘external’ data, i.e., table representation and chunks from the knowledge database; it is known that keeping external information at the end improves performance (Sui et al., 2024);
- Representing user tables in Markdown format, as LLMs are sensitive to different ways tabular inputs are serialized (Sui et al., 2024), and Markdown worked best for our use case (besides JSON and CSV).
- Providing descriptions of key column names and data types, e.g. “headers containing ‘goal’ always indicate booleans”, because headers are crucial to an LLM’s capacity to understand column contents and relations to other columns (Singha et al., 2023).
- Drawing on in-context learning (Fang et al., 2024), by describing three common topics in user queries (physical fatigue, mental fatigue, daily activities), the relevant columns in the user data. This ‘primes’ the LLM to consider relevant columns and possible relations between them.

Current work in prompt engineering research has further shown that, particularly in healthcare, a challenge is limited portability and scalability due to prompts becoming tied to specific domains, e.g., due to the use of terminology (Wang et al., 2025). Hence, we keep our approach straightforward, in that we leave specifics about the disease to the RAG component, and do not describe our medical population in detail. Our structuring and provision of examples in the prompt aim for portability, in that future users can easily adapt examples to their own data. In sum, our prompt is structured as follows:

1. A system message setting up the LLM as an AI assistant for addressing lifestyle-related questions for cancer patients;
2. Instructions for obtaining an empathic tone, brevity in the response, and refraining from providing medical advice;
3. A description of the user tables concerning headers (columns) and data types;
4. Several examples of question topics and the appropriate columns to consider;
5. Formatting instructions for separating user data analysis from counselling in the output;

¹See <https://healthychronos.com/en/> for more information.

²<https://github.com/yeem4n/llm-lifestyle-coach/>

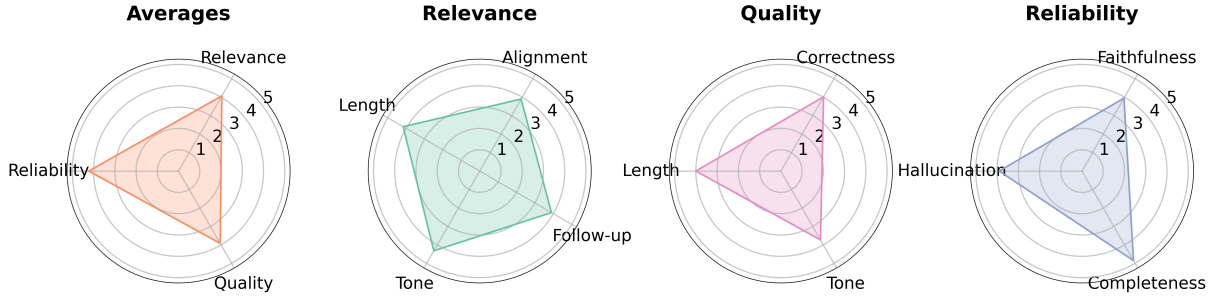


Figure 4: Visualization of QUORUM outcomes. ‘Averages’ graph summarises relevance, quality, and reliability; ‘Relevance’ and ‘Quality’ summarize their respective dimensions by averaging over variable scores; ‘Reliability’ contains rescaled variable ratios (Table 1.)

Var.	Var.	Var.
Alignm. 3.90 $\pm$.94	Correct 4.02 $\pm$.84	Faithfulness .79
Fol. up 3.90 $\pm$ 1.15	Tone 3.73 $\pm$.75	Completeness .97
Tone 4.32 $\pm$.82	Length 4 $\pm$.57	Hallucination .22
Length 4.16 $\pm$.69		
Users	Experts	Developers

Table 2: Main evaluation results. For users, we report averages over 19 users; for experts, averages over in total 100 evaluations (50 response-query pairs evaluated by two experts); for developers, ratios over in total 66 claims about user data (faithfulness/completeness) and 31 contextualisation statements (hallucination).

- Concatenation of table, RAG chunks, and user query to instructions above.

We used LangChain, an open-source framework to streamline LLM and RAG components (Chase, 2022), to build COACH.

The LLM used in the pipeline is GPT-4o (version 2024-11-20, accessed October 2025), due to workflows already in place for *Healthy Chronos*, in which they can easily deploy the system through the Azure environment, and its strong performance on similar tasks. GPT-4o is currently still a good option in similar, recent systems and chatbots for lifestyle counselling (Jörke et al., 2025; Steenstra et al., 2024); suitably large open models could in future work take the place of GPT-4o.

4. Results

4.1. User Evaluation: Relevance

Main results of applying QUORUM are given in Table 2 and visually in Figure 4. As can be seen, overall, users responded positively to the generated counselling. Results suggest that counselling aligned with personal circumstances, that tips were likely to be followed, and that the tone and length were satisfying, as all variable mean scores are around 4. However, we also see that variation in

scores is larger for statements on alignment and follow-up, implying that users are more divided on this aspect than on tone and length.

Remarks: Alignment Inspection of user-submitted remarks provided more insight into what users appreciate in the counselling. Ten out of nineteen respondents submitted an additional remark, of which four respondents expressed that the response captured their situation very well, and was easy to understand. They also appreciated the level of detail in the answers, and the formatting that keeps claims about the data separated from the contextualisation (Figure 1). In addition, two users emphasized that the advice felt more valuable due to the addition of specific references to their own diary data, which also helped them to better understand the contextualisation.

Remarks: Misalignment Four respondents expressed a misalignment in the counselling and their situation. One user, for example, reported the counselling recommended to seek help from friends and family, whereas for this user, social activities were experienced as tiring, and as an increasing burden for the social network, i.e., something to avoid. Other users mentioned a misalignment with their situation that is strictly speaking not related to the query-response pair, such as a lack of added functionality in the app, or struggling with regularly submitting data at the end of the day. One user noted that the advice reiterated strategies they were already applying without positive outcomes, implying the generated response ideally takes into account previous counselling in the prompt context.

Our user results suggest that the pipeline succeeds in generating advice that is relevant for users.

4.2. Expert Evaluation: Information Quality

Overall, experts were positive about the quality of generated counselling (Table 2, Figure 4). They

Annotator	A (n=25)	B (n=25)	C (n=12)	D (n=13)	E (n=12)	F (n=13)
Correctness	4.08 (± 0.63)	4.88 (± 0.33)	3.33 (± 0.47)	4.15 (± 0.36)	2.75 (± 0.73)	3.92 (± 0.47)
Tone	4.12 (± 0.59)	3.8 (± 0.63)	3.42 (± 0.76)	3.7 (± 0.92)	3 (± 0.71)	3.85 (± 0.36)
Length	4.16 (± 0.46)	4.04 (± 0.34)	3.92 (± 0.28)	4.46 (± 0.63)	3.25 (± 0.6)	3.92 (± 0.47)

Table 3: Mean and standard deviation of each expert and variable. N shows the number of query-response pairs evaluated per expert; each was labelled twice, totalling 100 observations.

thought the counselling was often correct, and had mostly the right tone and length, with most scores close to 4, but with smaller variation for each variable compared to users. Across 50 samples evaluated by 6 independent experts (minimal 2 evaluations per item with overlapping annotator pairs), Krippendorff’s α was 0.602 for correctness. This indicates moderate agreement: the experts generally concur about answer correctness. Examining the per-expert mean scores (Table 3), no average scores for correctness were found closer to two or one, suggesting disagreements reflect differences in annotator strictness rather than strong divergence about system correctness. Thus, the pipeline rarely produces counselling that experts consider incorrect or harmful.

Remarks: Tone Regarding tone, experts diverged. Krippendorff’s α was 0.150 for tone and 0.272 for length, confirming that these dimensions were very subjective. Inspection of a total of 79 remarks revealed that about a quarter raised concerns about tone and language use, describing the counselling as too condescending, direct, or unsympathetic. Also, the use of exclamation marks, imperatives, and motivational phrases was sometimes viewed negatively. Thus, experts see room for possible improvements, given that their evaluations were, on average, positive.

Remarks: Clarity Experts addressed clarity in about half of the remarks. Twenty-two remarks explicated that the counselling was not at a basic level (B1) of Dutch, hence should be rephrased. Experts considered some counselling statements on their own as too ambiguous, due to the lack of specific examples or further explanation, for example, the phrase ‘listen to your own body’ or ‘stay true to your routines’, as it is not always clear what this entails. Experts emphasized that clarity and concreteness in lifestyle advice are important, as this renders counselling more actionable and memorable, in line with previous research (Szymanski et al., 2025; Degachi et al., 2025).

Remarks: Grounding Five experts praised the explicit grounding of counselling in diary data, acknowledging the value of including insights from the user’s diary. While some did note that care should

be taken in not overgeneralizing over too little data, the references to concrete days, activities, or symptoms were seen as increasing transparency and trustworthiness.

Overall, though experts were quite positive about the counselling, they provided many remarks about how the counselling could be improved regarding its tone. In this respect, they were more critical than users, a finding which we return to in the discussion.

4.3. Developer Evaluation: Reliability

Overall, we found that 79% of claims about user patterns were faithful, i.e., consistent with the data. That is, about four out of five claims about user data are supported by the user’s table. This also means that specific referenced dates, activities, or goals from the user data are often correct, that is, concrete examples for patterns found in the user data. Further, regarding completeness, in 97% of the claims (virtually all), the LLM succeeded in finding the relevant information, which often amounted to correctly retrieving relevant columns and their values to address a user query. Lastly, regarding hallucination, in 22% of statements made in the contextualisation, we found information not strictly present in the external knowledge database.

Developer Agreement The developer agreement was overall moderate across faithfulness, completeness, and hallucination (Cohen’s $\kappa = 0.58$), indicating some challenges in evaluating claims and statements. In particular cases with more ambiguous phrases, e.g., ‘*The last few days...*’ or ‘*Scores are often between 3 and 5...*’ led to differing judgments, despite establishing a list with definitions for ambiguous terms (provided in the Supplementary Materials).

To better understand misalignment, we computed Positive Percentage Agreement for faithfulness, completeness, and RAG context adherence, which were 85.39%, 96%, and 88.89%, respectively, while Negative Percentage Agreement was 48%, 0%, and 44.44%, respectively. This indicates that the two developers often agree whenever one developer evaluates a claim as faithful, complete, or hallucination-free; however, developers have much more difficulty agreeing on the cases where, according to one of them, the LLM was unreliable,

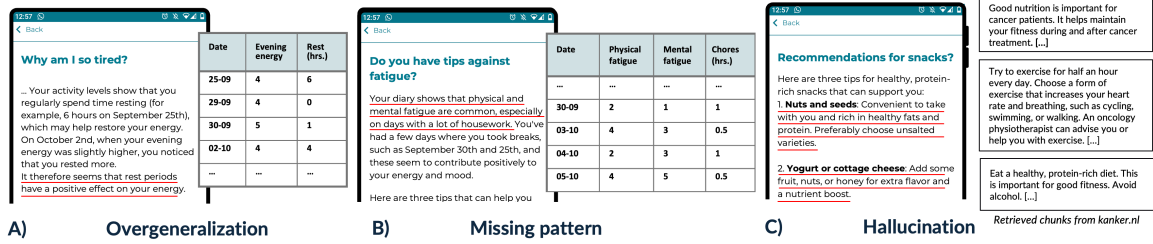


Figure 5: Examples of various errors in claims about data and contextualisation statements. A): overgeneralization of common-sense patterns: B) missing pattern in the user data; C): hallucination: advice is not strictly present in chunks.

incomplete, or provided information not present in the knowledge database.

Errors Though in most cases patterns in the data were extrapolated accurately, for faithfulness most errors arose from overgeneralisation, in which the model saw a pattern in the data that was not sufficiently strong or had also strong counterexamples. Example A in Figure 5 illustrates this. Another issue was with missing patterns: sometimes COACH reported on a pattern that did not exist (example B in Figure 5). This could be seen as a type of hallucination, which in QUORUM is captured under faithfulness.

Regarding hallucinations in contextualisation statements, we observed that the model sometimes goes beyond retrieved chunks in generating advice. Note that we had a strict interpretation of hallucination in Table 1: if a contextualisation statement suggested consuming healthy, protein-rich food, and the counselling contained examples (nuts, yoghurt, etc.) not strictly in the database, it was regarded as a hallucination. Arguably, this could also be seen as a case of intuitive extrapolation rather than harmful content often associated with hallucinations. When such cases (example C in Figure 5) were not counted as hallucination, the hallucination rate declined to 11%. Remaining errors contained reporting of insights from user data, contrary to the instruction to the LLM in the prompt. We found no cases of hallucinations with wrong or harmful content.

5. Discussion & Conclusion

This study contributed QUORUM, a framework for evaluating relevance, quality, and reliability of AI systems in health from three perspectives. The framework allows a future user to quickly glance at how overall, and in what specific aspects, a given AI-driven system in health scores on topics important to different stakeholders (Figure 4). As we have argued, a good evaluation of AI-driven systems in health should incorporate all perspectives in the

evaluation. We expect that AI systems that evaluate on average positively on relevance, quality, and reliability, which in practice will mean around 4 or higher, will have higher chances of making a positive real-world impact.

In our case, we saw convergence on a positive evaluation in the three perspectives, but also room for divergence. Divergence may reveal issues crucial for future development. In our case, we observed in expert remarks that they took issue with the clarity of the generated counselling. Interestingly, this did not seem to resonate with the scores or remarks of users, who indicated it was easy to follow. This signals our sample may be biased towards people with higher educational levels, as medical experts may have experience with a more varied population; this should inform the next iteration of COACH.

Some divergence can also be found in sensitivity to errors. Alignment and correctness as variables are arguably for users and experts proxies for developers' more technical faithfulness and completeness metrics (though they do not fully overlap). Still, the scores and remarks of users and experts do not suggest grave errors in user-specific pattern extraction or hallucinated content from the database, though we know from the developer evaluation that roughly one in five claims is not consistent with user data or is not strictly traceable to the knowledge database Table 2. This could be because different stakeholders have their own backgrounds and biases that may impact their sensitivity to errors; QUORUM helps reveal such differences as a starting point for further investigation.

We observed that if a user asks, for example, 'Why am I so tired?', the retrieved chunks may contain biomedical details, for example, on how changes in the immune system after cancer (therapy) can cause fatigue. Still, the counselling often defaults to more general tips, for example, on sleep patterns instead of more biomedical facts. This could be an unwanted result of prompting the model to refrain from providing medical advice; yet, the point here is that neither users/experts scores or remarks seem to pick up on this. Especially experts

could, due to their training, expect more comprehensive explanations in the counselling, which also includes more biomedical facts. A next iteration of COACH could focus on encouraging the LLM to include biomedical facts in the answer, which is not the same as providing medical advice.

In conclusion, this paper demonstrated that relatively straightforward LLM-based systems like COACH can provide relevant, good-quality, and reliable lifestyle counselling, as exemplified in our use case. We aimed to show that a multi-stakeholder lens is needed to evaluate the likelihood of the real-world impact of any AI system in health. We expect that both QUORUM and COACH, as an easy-to-understand framework and pipeline, will benefit other developers, experts, and patient populations aiming to improve consumer health question answering based on user diary data.

6. Ethical Statement

Our study protocol for recruitment of and data collection from users, medical experts, and developers was approved by the Leiden University Ethics Committee.

7. Acknowledgments

We thank the users of `Healthy Chronos` who contributed as participants to this study. We are grateful to the cancer information experts from `kanker.nl` for their time and effort in evaluating the responses. We also thank `Healthy Chronos` for the collaboration and access to the valuable platform.

8. Limitations

Several limitations impact our work. While we demonstrate the potential utility of our pipeline COACH and evaluation framework QUORUM for extension to arbitrary populations and diseases, these were only tested in a single context and use case, with a small population of users who likely have a high digital literacy. We hope to validate COACH and QUORUM across diverse settings and applications in future research.

Reproducibility is constrained for multiple reasons. Due to privacy and intellectual property considerations, the user diary data and the knowledge database cannot be shared. The RAG component naturally relies on dynamically evolving knowledge sources, which also limits the ability to reproduce our experimental setup entirely. Nevertheless, we anticipate that the components of our pipeline can be adapted and reused in other contexts where user diary data is available.

In addition, our sample size of evaluated responses was relatively small. The sample size and lack of baseline comparisons were primarily driven by the limited availability of our evaluators. More research on automatic lifestyle counselling using the framework and pipeline would help strengthen our conclusions.

Furthermore, the LLM-generated counselling was based only on the most recent <21 days of diary data. Future research could investigate how performance changes when incorporating longer-term lifestyle patterns. Finally, our pipeline relies on a single proprietary LLM (GPT-4o), as discussed in [subsection 3.2](#). Future work should explore the use of open-source, and ideally smaller, models to improve accessibility, transparency, and reproducibility.

9. Bibliographical References

Tatsuo Akechi, Akira Kugaya, Hitoshi Okamura, Shigeto Yamawaki, and Yosuke Uchitomi. 1999. Fatigue and its associated factors in ambulatory cancer patients: a preliminary study. *Journal of pain and symptom management*, 17(1):42–48.

Alexander Bondarenko, Ekaterina Shirshakova, Marina Driker, Matthias Hagen, and Pavel Braslavski. 2021. [Misbeliefs and biases in health-related searches](#). In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 2894–2899, New York, NY, USA. Association for Computing Machinery.

Nattapat Boonprakong. 2025. *Quantifying and Understanding the Effects of Cognitive Biases in Human-Computer Interaction*. Ph.D. thesis, The University of Melbourne, Melbourne, Australia.

Harrison Chase. 2022. [Langchain](#). Accessed: 2025-11-23.

Chadha Degachi, Ujjayan Dhar, Evangelos Niforatos, and Gerd Kortuem. 2025. [Towards a Domain Expert Evaluation Framework for Conversational Search in Healthcare](#). In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Rachel Drbohlav Ollerton, Saar Hommes, Nadine Bol, Gwenn Beets, Frans Folkvord, and Emiel J. Krahmer. 2025. [Health-related lifestyle communication in oncology: Exploring the experiences and preferences of cancer survivors](#). *Health Communication*, page 1–10.

- Xi Fang, Weijie Xu, Fiona Anting Tan, Ziqing Hu, Jiani Zhang, Yanjun Qi, Srinivasan H. Sengamedu, and Christos Faloutsos. 2024. [Large language models \(LLMs\) on tabular data: Prediction, generation, and understanding - a survey](#). *Transactions on Machine Learning Research*.
- Meeke Hoedjes, Inge Nijman, and Chris Hinnen. 2022. Psychosocial determinants of lifestyle change after a cancer diagnosis: a systematic review of the literature. *Cancers*, 14(8):2026.
- Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A. Landay. 2025. [Gptcoach: Towards llm-based physical activity coaching](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.
- Lars Kayser, Andre Kushniruk, Richard H Osborne, Ole Norgaard, and Paul Turner. 2015. Enhancing the effectiveness of consumer-focused health information technology systems through ehealth literacy: a framework for understanding users' needs. *JMIR human factors*, 2(1):e3696.
- Xiangxun Lai, Yue Lai, JiaCheng Chen, Shengqi Huang, Qi Gao, and Caihua Huang. 2025. Evaluation strategies for large language model-based models in exercise and health coaching: Scoping review. *Journal of Medical Internet Research*, 27:e79217.
- 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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.
- Karen Jia-Hui Li, Simone Balloccu, Ondrej Dusek, and Ehud Reiter. 2025. [When LLMs can't help: Real-world evaluation of LLMs in nutrition](#). In *Proceedings of the 18th International Natural Language Generation Conference*, pages 753–779, Hanoi, Vietnam. Association for Computational Linguistics.
- Mike A Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, et al. 2024. Transforming wearable data into health insights using large language model agents. *arXiv preprint arXiv:2406.06464*.
- Sophia Meywirth. 2024. Designing a large language model-based coaching intervention for lifestyle behavior change. In *International Conference on Design Science Research in Information Systems and Technology*, pages 81–94. Springer.
- Siddharth Nandagopal. 2025. [Tokens as currency: A novel framework to sustain ai adoption and profitability](#). *International Journal of Applied Information Systems*, 12(46):44–51.
- Diana Bar-Or Nirman, Ariel Weizman, and Amos Azaria. 2025. Fool me, fool me: User attitudes toward llm falsehoods. *IEEE Access*.
- George Papageorgiou, Vangelis Sarlis, Manolis Maragoudakis, Ioannis Magnisalis, and Christos Tjortjis. 2025. Evaluating faithfulness in agentic rag systems for e-governance applications using llm-based judging frameworks. *Big Data and Cognitive Computing*, 9(12):309.
- Alexander Pugachev, Ekaterina Artemova, Alexander Bondarenko, and Pavel Braslavski. 2023. Consumer health question answering using off-the-shelf components. In *Advances in Information Retrieval*, page 571–579. Springer Nature Switzerland.
- Rebecca Raeside, Si Si Jia, Julie Redfern, and Stephanie R Partridge. 2022. Navigating the online world of lifestyle health information: qualitative study with adolescents. *JMIR Pediatrics and Parenting*, 5(1):e35165.
- Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. [Ai and the everything in the whole wide world benchmark](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Ananya Singha, José Cambronero, Sumit Gulwani, Vu Le, and Chris Parnin. 2023. [Tabular representation, noisy operators, and impacts on table structure understanding tasks in LLMs](#). In *NeurIPS 2023 Second Table Representation Learning Workshop*.
- Ian Steenstra, Farnaz Nouraei, Mehdi Arjmand, and Timothy Bickmore. 2024. Virtual agents for alcohol use counseling: Exploring llm-powered motivational interviewing. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, pages 1–10.
- 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, Merida Mexico. ACM.

- 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, IUI '25*, pages 952–966, New York, NY, USA. Association for Computing Machinery.
- Marrit Annika Tuinman, Janine Nuver, Anke De Boer, Anne Looijmans, and Mariët Hagedoorn. 2024. [Lifestyle changes after cancer treatment in patients and their partners: a qualitative study](#). *Supportive Care in Cancer*, 32(4):248.
- Bente van Aken, Anna Manshanden, Willemieke Kroeze, Nadine Florisson, Barbara Groot-Sluijsmans, Meeke Hoedjes, Kristel van Asselt, Lies ter Beek, and Ingrid Steenhuis. 2025. [Experiences and needs of dutch cancer survivors regarding lifestyle counselling: a qualitative study](#). *BMC Cancer*, 25:1761.
- Bram M.A. Van Dijk, Armel E.J.L. Lefebvre, and Marco R. Spruit. 2025. [Welzijn.ai: Developing responsible conversational ai for the care of older people through stakeholder involvement](#). *Maturitas*, 199(108616).
- Julia EWC van Gemert-Pijnen, Nicol Nijland, Maarten Van Limburg, Hans C Ossebaard, Saskia M Kelders, Gunther Eysenbach, and Erwin R Seydel. 2011. A holistic framework to improve the uptake and impact of ehealth technologies. *Journal of medical Internet research*, 13(4):e1672.
- Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Huawen Hu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, et al. 2025. Prompt engineering for healthcare: Methodologies and applications. *Meta-Radiology*, page 100190.
- Jiayin Wang, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024a. Understanding user experience in large language model interactions. *arXiv preprint arXiv:2401.08329*.
- Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024b. [A user-centric multi-intent benchmark for evaluating large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3612, Miami, Florida, USA. Association for Computational Linguistics.

Addressing Domain Shift in Health Coaching Note Analysis through Factorized Synthetic Data Generation

Michael Tänzer^{1,2}, Iva Bojic¹, Ashwini Lawate¹, Andy Hau Yan Ho,¹
Andy W. H. Khong¹

¹ Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

² Imperial College London, Imperial Global Singapore, Singapore

{cs-michael.tanzer, iva.bojic, ashwini.lawate, andyhyho, andy.khong}@ntu.edu.sg

Abstract

Automatic extraction of behavioral goals from health coaching notes is essential for scalable monitoring of coaching programs, yet training data is scarce and exhibits substantial domain shift across programs. We collect and annotate 157 notes from a coaching program and show that models trained on the only existing public corpus, SMARTSpan (173 notes), suffer a drop of up to 30 percentage points in exact-match F1 when transferred to our data. To address this, we propose a factorized synthetic data generation pipeline that decomposes note variation into three largely independent axes (health coach documentation structure, patient goal content, and patient persona), extracts empirical priors from a small in-domain seed set, and samples from them to generate diverse synthetic notes with embedded goal-span labels validated via cycle-consistency filtering. In low-resource experiments with only 57 in-domain training notes, our approach outperforms rephrasing and back-translation baselines on both exact-match and partial-match F1. Ablation analysis demonstrates that augmentation must target the in-domain distribution to be effective, and a human evaluation confirms that synthetic notes are structurally faithful, with detection driven by surface artifacts rather than content or organizational flaws. All code and generated data will be published at the following GitHub repository: [cl4health-factorized-augmentation](https://github.com/cl4health-factorized-augmentation).

Keywords: health coaching, synthetic data generation, domain adaptation, goal span extraction, low-resource NLP, data augmentation

1. Introduction

Health coaching is a person-centered intervention that supports prevention, self-management, and sustained behavior change in individuals at risk of chronic disease (Wolever et al., 2013; Loughnane et al., 2025), with demonstrated benefits for physical activity, cardiovascular risk, and quality of life (Olsen and Nesbitt, 2010; Kivelä et al., 2014). Central to behavioral coaching is the formulation of Specific, Measurable, Attainable, Relevant, and Time-bound (SMART) goals, which support clients in maintaining momentum between sessions and adhering to agreed actions (Doran, 1981; Wallace et al., 2018; White et al., 2020; Bahrami et al., 2022). Health coaches record these goals in free-text session notes (Gupta et al., 2021; Bojic et al., 2025), making automatic extraction of goal spans a prerequisite for scalable monitoring and evaluation (Flocke and Stange, 2004; Bowman et al., 2015; Zhou et al., 2024; Loughnane et al., 2025).

Training span extraction models for this task is hampered by the lack of suitable annotated data. Conversational and counseling corpora emphasize emotional support or dialogue structure rather than measurable behavior change (Malhotra et al., 2022; Xu et al., 2025; Qi et al., 2025), and resources that do include goal-related content are often small, narrowly scoped, or limited to short text-based interactions (Gupta et al., 2020, 2021; Zhou et al.,

2024). The closest publicly available resource, SMARTSpan (Bojic et al., 2025), provides 173 annotated notes from a single randomized controlled trial targeting cardiovascular risk reduction through statin adherence and lifestyle modification among patients with hyperlipidemia. To assess transferability, we collect and annotate a new corpus of 157 notes from 56 clients seen by three health coaches in a separate coaching program. Retraining the SpanQualifier-based (Huang et al., 2023) benchmark models from SMARTSpan (Bojic et al., 2025) and evaluating them on our corpus reveals a sharp performance drop (Table 1), and a t-distributed Stochastic Neighbor Embedding (t-SNE; van der Maaten and Hinton 2008) projection of note embeddings confirms that the two corpora occupy distinct regions of the representation space (Figure 1).

Given the cost of full-scale annotation, a natural alternative is to augment limited in-domain data with synthetic examples. However, surface-level augmentation methods such as paraphrasing or back-translation are insufficient because variation in health coaching notes is not purely lexical—such variation arises from the interaction of health coach-specific documentation *structure*, patient-specific *goal content*, and *persona* context. We propose a corpus-grounded generation pipeline that factorizes these axes, extracts empirical priors from a small set of in-domain notes, and samples from them to produce diverse synthetic notes with em-

bedded goal-span labels. A cycle-consistency filter ensures label quality before the synthetic data is used for training.

The contributions of this work are as follows:

1. **Factorized Augmentation Pipeline:** We propose a corpus-grounded generation pipeline that decomposes health coaching note variation into structure, goal content, and persona axes before extracting empirical priors from a small in-domain corpus, and sampling from them to produce diverse synthetic notes with embedded goal-span labels validated via cycle-consistency filtering.
2. **Comprehensive Low-Resource Evaluation:** We conduct systematic experiments across multiple training regimes and ablation settings, demonstrating that factorized generation consistently outperforms general-purpose augmentation baselines and recovers strong span extraction performance from limited in-domain data.
3. **In-Domain Enrichment:** We demonstrate that factorized generation also improves performance when applied within an established corpus, indicating that the pipeline provides value beyond cross-domain adaptation by enriching the training signal even in well-resourced single-domain settings.

2. Method

2.1. Data

Our dataset comprises 157 health coaching sessions collected from 56 unique clients and delivered by three professional health coaches. All notes are written in English. Sessions were conducted as part of a structured health coaching program involving education, medication adherence support, and behavioral goal setting, with clients participating in repeated follow ups over time (mean = 2.82 ± 1.73 sessions per client). Among the 55 clients with available demographic information, 28 (50.9%) were male. The majority were of Chinese ethnicity ($n = 43$, 78.2%), followed by Malay ($n = 5$, 9.1%), Indian ($n = 3$, 5.5%), and other ethnic backgrounds ($n = 4$, 7.3%). Participants were middle-aged to older adults, ranging from 22 to 77 years (mean = 57.3 ± 10.3), reflecting a population typically targeted for long-term lifestyle modification and chronic disease prevention interventions. Table 2 summarizes key characteristics of our dataset in comparison with SMARTSpan, highlighting differences in note length, goal density, and overall structure.

Our 157 notes were annotated for goal spans following guidelines aligned with those of Bojic et al.

(2025) including, for instance, removing leading function words such as *to* from span boundaries. The notes were randomly partitioned into three equal subsets and distributed among three annotators such that each note was independently labeled by exactly two annotators. The annotators then met jointly to review all notes on which any disagreement had been identified (47 of 157 notes, affecting 175 of 509 goal-span comparisons) and resolved for each case through discussion until full consensus was reached. Inter-annotator agreement was measured on the dual-annotated notes prior to adjudication: word-level Cohen’s $\kappa = 0.94$ overall (0.92 - 0.96 across annotator pairs), and goal-level $F1 = 0.79$ (0.77 - 0.81), indicating substantial agreement on both token-level span boundaries and goal identification. The adjudicated labels serve as the gold standard for all experiments.

2.2. Data Generation

Health coaching notes exhibit variation along three largely independent axes: the *structure and style* imposed by the health coach’s documentation habits, the *semantic content* determined by the patient’s agreed goals, and the *persona* reflecting the patient’s demographic, personal preference, and contextual profile. Surface-level augmentation methods such as paraphrasing and back-translation operate only on the lexical form of existing notes and cannot introduce new patient personas, vary the structural patterns that characterise different coaches, or alter the distribution of goal attributes.

We therefore propose a corpus-grounded generation pipeline that factorizes these axes, extracts empirical priors from the actual corpus, and samples from them to produce diverse, labeled synthetic notes. The pipeline proceeds in two phases: a *corpus analysis* phase that extracts distributional priors along each axis of variation and a *generation* phase that samples from those priors to produce synthetic notes with embedded goal-span labels. An overview of the above process is illustrated in Figure 2.

2.2.1. Corpus Analysis

The corpus analysis phase performs three passes over the real notes, one per axis of variation, to extract the distributional priors that later guide synthetic note generation. All extraction calls use Gemini 3.0 Pro with temperature 0 to ensure deterministic outputs; the full prompts and output schemas are provided in Appendix B.

Goal Attribute Enrichment. Each annotated goal span within the corpus is processed individually to extract a structured set of attributes grounded

Model	SMARTSpan dataset		Our dataset	
	EM F1	PM F1	EM F1	PM F1
BERT-large-cased	90.21 $\pm$ 2.29	94.35 $\pm$ 2.80	62.39 $\pm$ 2.15	74.79 $\pm$ 3.41
DeBERTa-v3-large	97.56 $\pm$ 2.03	98.27 $\pm$ 1.34	71.85 $\pm$ 2.38	81.78 $\pm$ 3.39

Table 1: Cross-domain evaluation of models trained on SMARTSpan and tested on both datasets. Exact-Match F1 (EM F1) requires predicted and gold spans to match exactly; Partial-Match F1 (PM F1) awards credit proportional to token overlap. Results report mean $\pm$ standard deviation over five random data splits.

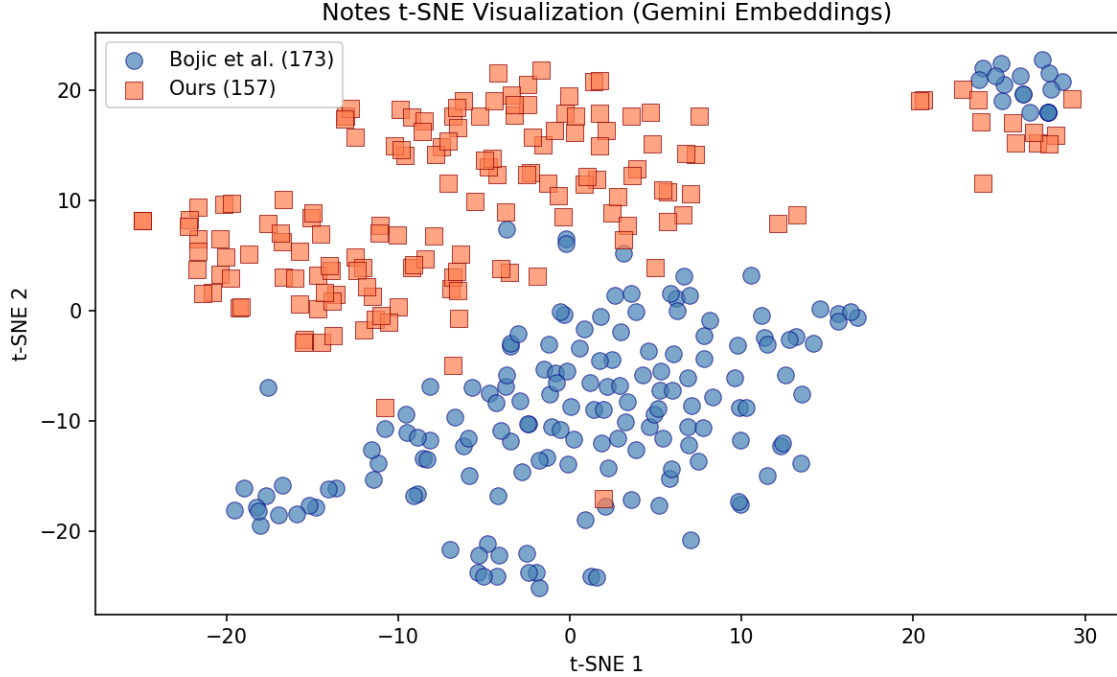


Figure 1: t-SNE visualization of embeddings comparing the SMARTSpan dataset (Bojic et al., 2025) ($n = 173$) and our collected dataset ($n = 157$). The distinct clustering highlights a significant domain shift that explains the performance degradation of the baseline models.

	SMARTSpan	Our dataset
# notes	173	157
# health coaches	3	3
# goals	266	331
Goals per note (avg.)	1.5	2.1
Max goals per note	5	7
Goals per note (90th pct.)	3	4
Zero-goal notes	46 (26.6%)	41 (26.1%)
Words per note (avg.)	161	469
Characters per note (avg.)	1,052	2,678

Table 2: Dataset summary statistics for the SMARTSpan and our datasets.

in the behavioral goal taxonomy of Hessler et al. (2019). The whole note is used as context for each goal’s extraction. Each span is assigned a goal type from the five categories listed in Table 3, together with the specific activity or behavior referenced and, where present, optional attributes: location, time

Goal Type	Description
Physical activity	Exercise or movement goals
Dietary behavior	Nutrition and eating goals
Med. adherence	Taking prescribed medication
Substance use	Reducing substance use
Self-monitoring	Tracking health behaviors

Table 3: Goal type categories, following the taxonomy of Hessler et al. (2019).

of day, frequency (value and unit), duration (value and unit), timeframe (type and text), conditionality, and patient confidence (type and value). Notes containing no goals receive a separate classification by reason (Table 4). Corpus-wide frequency distributions over all attribute values form the prior from which goal configurations are sampled during generation.

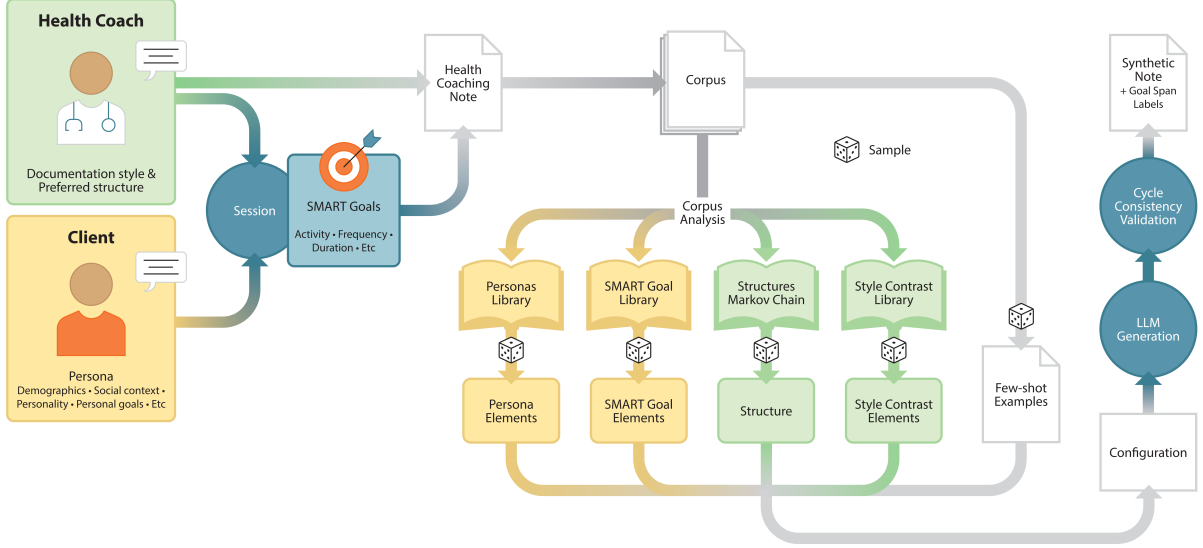


Figure 2: Overview of the factorized synthetic data generation pipeline. The process begins with a corpus analysis phase that decomposes real health coaching notes into independent axes of variation: client persona, SMART goal content, documentation structure, and stylistic contrast. Empirical priors are extracted from the corpus to populate respective libraries. During the generation phase, elements are independently sampled from these libraries to assemble a target configuration. A Large Language Model uses this configuration, alongside few-shot examples drawn from the corpus, to produce a synthetic note with embedded goal span labels. Finally, a cycle-consistency validation step filters the output to ensure label quality.

Reason	Description
No-show	Patient did not attend the session
Declined	Patient declined to set goals
Review-only	Session focused on reviewing goals
Other	Other administrative reasons

Table 4: Zero-goal note classifications.

Structural Analysis. To model the health coach structure axis, each note is decomposed into an ordered sequence of functional sections. Section types are drawn from a closed vocabulary of nine categories defined in consultation with certified health coaches based on the session checklist used in clinical practice: weekly goals review, long-term goals review, app usage review, journaling review, goal setting, barriers discussion, patient information, generative moment (collaborative exploration of values or new directions), and other. Each section is additionally labeled with its format (numbered list, unordered list, or embedded narrative), narrative voice (third person, first person, coach-we, passive, or mixed), and dominant tense pattern (primarily past, primarily future, or mixed). Post-processing removes consecutive duplicate section types.

From the resulting sequences, we construct a first-order Markov chain over section types to model how health coaches habitually organise their notes.

Let $\mathcal{S} = \{s_1, \dots, s_9\}$ denote the nine section types and let END denote a terminal state. The chain comprises an initial-state distribution $\pi_0(s) = P(S_1 = s)$, a transition matrix $P(s' | s)$ for $s, s' \in \mathcal{S} \cup \{\text{END}\}$, and an empirical sequence-length distribution $P_L(\ell)$ estimated from the observed number of sections per note. Transition probabilities are smoothed using interpolated Kneser-Ney smoothing (Kneser and Ney, 1995) with a discount factor $d = 0.75$. At generation time, the target sequence length ℓ is first drawn from P_L , then the chain is forward-stepped for exactly ℓ transitions starting from a section type sampled from π_0 .

Persona Extraction. To model the patient axis beyond goal content, each note is processed to extract a structured set of persona elements derived from commonly used definitions of sociodemographic factors (National Library of Medicine, 2024) and Social Determinants of Health as defined by the World Health Organization (World Health Organization, 2019). Social demographics are extracted as structured fields: age group, gender, ethnicity, occupation, education level, marital status, and income level. Four categories of contextual elements are extracted as free-text lists using snake_case normalization to facilitate deduplication and frequency counting: *social context* (e.g. `lives_with_partner`), *economic context* (e.g. `late_shifts`), *physical environment* (e.g.

access_to_gym), and *individual characteristics* (e.g. routine_dependent). Results are aggregated into a persona library recording unique element sets and corpus-wide frequency distributions. The compositional structure of this library allows new patient descriptions to be assembled element-by-element rather than copied from any real record, supporting both diversity and privacy preservation.

2.2.2. Style Contrast Library

Health coach-level stylistic variation is captured through a library of paired real and synthetic notes ranked by stylistic divergence. A diverse subset of real notes is selected from the corpus via stratified round-robin sampling over (*health coach*, *goal count bucket*) strata, maximizing coverage across coaches and session types. For each selected note, a synthetic twin is generated by Gemini 3.0 Pro ($T=1.0$) using a configuration constructed directly from that note’s own extracted attributes (goals, structure, persona) rather than sampled from corpus distributions. This grounds each twin to its real counterpart, so that embedding distance between the pair predominantly reflects stylistic divergence rather than content divergence. We note that information lost during attribute extraction introduces some residual content variation; however, this information bottleneck is consistent across all pairs, so the resulting noise is approximately uniform and does not systematically bias distance-weighted sampling toward content-divergent pairs.

Both notes are embedded using Gemini Embedding (Lee et al., 2025) and cosine distance is computed per pair. During generation, a style contrast pair is selected from the library with probability proportional to its cosine distance, providing the generation model with a concrete positive (real) and negative (synthetic) stylistic anchor.

2.2.3. Synthetic Note Generation

Each synthetic note is produced through a structured sampling and prompting procedure.

Configuration Sampling. For each note to generate, a full configuration is assembled by sampling independently from the corpus-derived priors along each axis.

Structure. The target number of sections ℓ is drawn from the empirical length distribution P_L . An initial section type is sampled from π_0 , and the chain is forward-stepped for ℓ transitions using the Kneser-Ney smoothed transition matrix.

Goals. The number of goals is sampled from the corpus distribution of goal counts. For each goal, attributes are sampled sequentially: goal type from the marginal corpus distribution, activity from the type-conditioned activity list, and each optional

attribute a (frequency, duration, timeframe, conditionality, confidence) sampled with null probability

$$P(a = \emptyset) = \hat{p}_{\emptyset}^{(a)} = \frac{|\{g \in \mathcal{G} : a_g \text{ is absent}\}|}{|\mathcal{G}|}, \quad (1)$$

where $\mathcal{G}$ is the set of all annotated goals in the corpus. Conditional on $a \neq \emptyset$, the value is drawn from the empirical distribution of non-null values for that attribute, preserving the realistic sparsity observed in the corpus.

Persona. Social demographic fields are sampled from corpus distributions, each with a per-field null probability estimated analogously to Equation 1. List-type persona categories are populated by first drawing a count from the empirical count distribution for that category, then sampling that many elements without replacement from the category-specific element library.

Few-shot examples. Five few-shot examples are drawn from the corpus using a tiered filter relaxation strategy that prioritizes stylistic similarity while ensuring availability: selection first attempts to match on health coach identity, then relaxes successively to note format, goal count, zero-goal reason, and finally random selection.

Style contrast. A single style contrast pair is injected into the prompt, selected from the precomputed library with probability proportional to the cosine distance between the real note and the synthetic note.

Prompt Construction and Generation. The sampled configuration is assembled into a structured prompt and passed to Gemini 3.0 Pro ($T=1.0$). Separate templates are used for goal-bearing and zero-goal notes; the latter omits persona and structure sections, which are largely absent from no-show or administrative records. The model returns the full note text together with goal spans as exact substrings in a single structured output call, eliminating the need for post-hoc span alignment.

2.2.4. Cycle-Consistency Validation

Generated notes are filtered for label quality using a cycle-consistency criterion inspired by the forward-backward consistency checks common in unsupervised learning (Zhu et al., 2017). A separate Gemini 3.0 Pro call independently re-extracts goal spans from each generated note with no access to the generation labels. Re-extracted spans are compared against the generated labels using token-level F1: for each generated goal span g_i and its best-matching re-extracted span $\hat{g}_i$, we compute

$$F1(g_i, \hat{g}_i) = \frac{2 |\text{tok}(g_i) \cap \text{tok}(\hat{g}_i)|}{|\text{tok}(g_i)| + |\text{tok}(\hat{g}_i)|}, \quad (2)$$

where $\text{tok}(\cdot)$ denotes the token set of a span. The note-level score is the mean F1 across all generated goals.

To account for noise in the re-extraction model itself rather than genuine label errors, up to three re-extraction trials are performed at $T=0.7$ and the best score across trials is retained. A note is accepted only if this best score reaches a threshold of 1.0, i.e. at least one re-extraction trial recovers all generated labels exactly at the token level, without adding or removing goals or text in any of the goals. This strict criterion reflects the fact that training downstream models on partially incorrect span labels introduces systematic noise into the learned representations.

Under this protocol, approximately 20% of generated notes are rejected. Failed notes are saved with their validation metadata and can be rechecked under revised configurations without discarding already-accepted examples. The generation loop continues until a target number of accepted notes is reached.

2.3. Experimental Setting

Data Splits. From our 157-note corpus we randomly sample 57 notes for training and reserve the remaining 100 as a held-out test set. The 57-note training set is deliberately small to simulate the realistic low-resource scenario in which a new health coaching program (or an existing program undergoing severe distribution shift after, for instance, onboarding several new health coaches) can only afford limited annotation effort. The full SMARTSpan corpus (173 notes) serves as the out-of-domain training resource. For experiments evaluated on the SMARTSpan test set (Table 6), we follow a 60/20/20 train/validation/test split over the 173 notes, repeated five times with different random partitions. Synthetic notes whose source examples overlap with the test fold are excluded from training and validation to prevent data leakage.

Augmentation Conditions. We compare three augmentation strategies applied to the 57 in-domain notes:

- **Rephrasing:** each note is rewritten by Gemini 3.0 Pro with instructions to paraphrase the text while jointly paraphrasing and extracting the goal spans.
- **Backtranslation:** each note is translated from English to a language selected at random from a list comprised of all languages that appear in more than 1% of the available text on the internet (Statista, 2025), and back using Gemini 3.0 Pro, with a post-processing step that realigns goal-span boundaries to the backtranslated text via fuzzy string matching.

- **Ours (factorized generation):** synthetic notes are generated following the pipeline described in Section 2.2, using the 57 in-domain notes as the analysis corpus.

Each method produces 200 synthetic notes per condition. When augmentation is applied to the SMARTSpan corpus, the same procedure is used with SMARTSpan as the analysis corpus. Figure 3 provides an overview of the dataset composition across all experimental conditions.

Training Configurations. Table 5 reports results under two regimes. In the *low-resource* regime, only the 57 in-domain notes (with or without augmentation) are used for training. In the *combined* regime, the full SMARTSpan corpus is concatenated with the 57 in-domain notes before augmentation is applied to the in-domain subset. An additional *cross-domain baseline* trains exclusively on SMARTSpan and evaluates on our test set, quantifying the domain shift reported in Introduction.

Models and Training. Following Bojic et al. (2025), we fine-tune two pre-trained encoders: BERT-large-cased (Devlin et al., 2019) and DeBERTa-v3-base (He et al., 2020) with a token classification head for BIO-tagged goal span extraction. All models were trained for 100 epochs using the AdamW optimizer (Loshchilov and Hutter, 2017), with a learning rate of 3×10^{-5} , a training batch size of 32, an evaluation batch size of 1, and gradient accumulation over 4 steps. Input sequences were truncated to a maximum length of 512 tokens, and experiments were conducted with a fixed random seed of 30.

Evaluation Metrics. We report two span-level F1 scores: *Exact-Match F1* (EM F1), which requires predicted and gold spans to match exactly, and *Partial-Match F1* (PM F1), which awards credit proportional to the token overlap between predicted and gold spans Li et al. (2022). Both metrics are computed at the note level and averaged across the test set.

Statistical Protocol. All experiments are repeated over five random splits to account for variance due to data partitioning. We report the mean and standard deviation of each metric across the five runs as *mean $\pm$ std*.

Human Distinguishability Test. To assess the perceptual realism of the generated notes, we conducted a two-alternative forced-choice experiment in which human evaluators were asked to distinguish real notes from synthetic ones. In each trial, one genuine note sampled from our corpus and

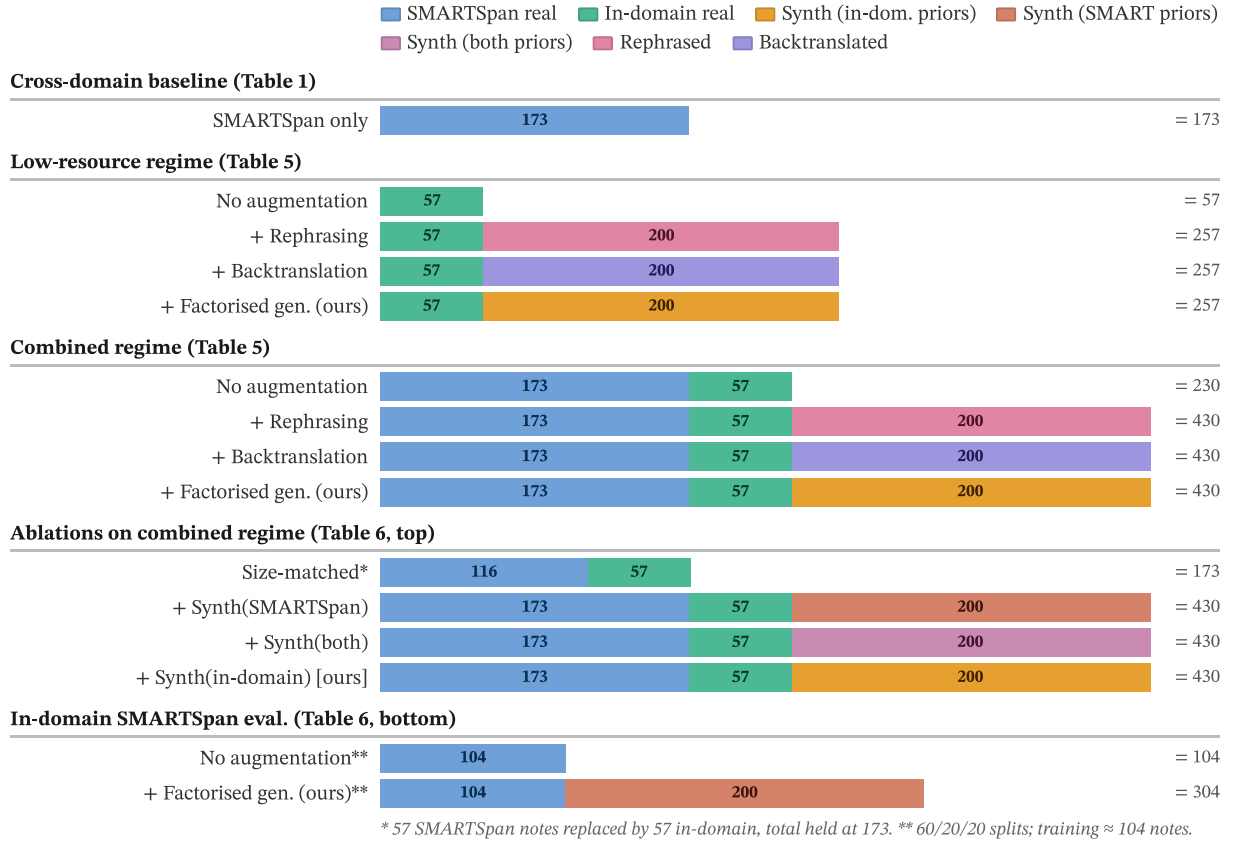


Figure 3: Composition of training sets across all experimental conditions. Each bar shows the number and source of notes used for training. Synth(in-domain) and Synth(s) denote synthetic notes generated from in-domain and SMARTSpan priors, respectively.

one synthetic note produced by the factorized generation pipeline were presented side by side in random order, and the evaluator was asked to select the note they believed to be real. Two evaluators, both familiar with health coaching documentation, completed a total of 50 trials (25 each). We report detection accuracy as the proportion of trials in which the evaluator correctly identified the real note, with chance performance at 50%.

3. Results

Domain shift confirms the need for adaptation. Training exclusively on SMARTSpan and evaluating on our corpus yields EM F1 scores of 62.39 (BERT) and 71.85 (DeBERTa), well below the in-domain performance reported by Bojic et al. (2025). This cross-domain gap, consistent with the distributional divergence visualized in Figure 1, motivates all subsequent experiments.

Factorized generation is the strongest augmentation in low-resource settings. With only 57 in-domain notes and no external data, our pipeline lifts EM F1 from 75.28 to 80.85 (+5.57) for BERT and from 82.79 to 87.30 (+4.51) for DeBERTa (Table 5,

low-resource group). Both surface-level baselines also improve over the unaugmented condition, but neither matches factorized generation: backtranslation reaches 79.24 and 80.49, while rephrasing reaches 78.02 and 83.39 for BERT and DeBERTa, respectively. The advantage is most pronounced on PM F1, where our method achieves 90.01 and 96.08 compared to 87.16 and 91.60 without augmentation, suggesting that the factorized sampling introduces greater diversity in span boundaries.

Augmentation benefits persist when out-of-domain data is available. Adding SMARTSpan to the 57 in-domain notes (Table 5, combined group) raises unaugmented EM F1 to 79.99 (BERT) and 84.13 (DeBERTa), confirming that out-of-domain data provides a useful prior despite the domain shift. Applying factorized generation to the in-domain subset further improves performance to 82.38 and 85.01, again outperforming both backtranslation (82.29, 84.89) and rephrasing (78.99, 82.62). Notably, rephrasing slightly degrades performance relative to the unaugmented combined baseline for both models, likely because lexical variation alone cannot compensate for the additional noise introduced by imperfect span realignment.

Training data	BERT		DeBERTa	
	EM F1	PM F1	EM F1	PM F1
<i>Cross-domain baseline</i>				
SMARTSpan only	62.39±2.15	74.79±3.41	71.85±2.38	81.78±3.39
<i>Low-resource (57 in-domain notes)</i>				
No augmentation	75.28±3.26	87.16±0.91	82.79±3.81	91.60±3.83
+ Rephrasing	78.02±1.65	85.66±1.77	83.39±4.36	91.24±1.84
+ Backtranslation	79.24±1.89	87.16±2.24	80.49±2.34	90.52±2.64
+ Factorised generation (ours)	80.85±1.85	90.01±1.38	87.30±1.27	96.08±1.48
<i>Combined (SMARTSpan + 57 in-domain notes)</i>				
No augmentation	79.99±2.90	90.41±2.25	84.13±0.88	93.56±0.79
+ Rephrasing	78.99±1.91	87.89±2.14	82.62±2.64	92.62±1.52
+ Backtranslation	82.29±1.91	90.77±1.81	84.89±2.52	94.41±2.10
+ Factorised generation (ours)	82.38±1.33	91.45±1.74	85.01±1.78	94.14±1.43

Table 5: Goal span extraction performance on our held-out test set ($n = 100$). All models are evaluated on the same 100 in-domain test notes. Results report mean $\pm$ standard deviation over five random data splits. Best result per column within each group is reported in **bold**.

Configuration	BERT		DeBERTa	
	EM F1	PM F1	EM F1	PM F1
<i>Ablations (tested on our held-out set)</i>				
SMARTSpan + 57 ID (size-matched)*	81.31±3.81	90.94±1.10	84.76±2.34	94.08±1.40
SMARTSpan + Synth(s) [†] + 57 ID	77.86±3.86	89.63±2.05	82.13±2.62	93.12±1.33
SMARTSpan + 57 ID + Synth(both) [‡]	80.88±1.44	91.14±1.19	84.49±1.44	93.82±2.14
SMARTSpan + 57 ID + Synth(ID) [ours]	82.38±1.33	91.45±1.74	85.01±1.78	94.14±1.43
<i>In-domain SMARTSpan (tested on SMARTSpan)</i>				
No augmentation	90.21±2.29	94.35±2.80	97.56±2.03	98.27±1.34
+ Factorised generation (ours)	91.44±3.00	95.70±2.22	97.08±2.01	98.43±1.50

* 57 SMARTSpan notes replaced with 57 ID notes (total training set size held constant).

[†] Synth(s) = factorised generation applied to SMARTSpan notes only. [‡] Synth(ID) = applied to the 57 ID notes; Synth(both) = applied to both corpora.

Table 6: Ablation studies and in-domain (ID) SMARTSpan evaluation. *Top*: variations on the combined training regime, tested on our held-out set ($n = 100$). *Bottom*: models trained and tested on SMARTSpan using five random 60/20/20 splits, with synthetic examples whose source IDs overlap the test fold excluded from training. Results report mean $\pm$ standard deviation. Best result per column within each group is reported in **bold**. ID = in-domain.

Augmenting out-of-domain data does not help.

The ablation study in Table 6 reveals that applying factorized generation to the SMARTSpan corpus rather than the in-domain notes (“Synth(s) + 57 in-domain”) hurts EM F1 by 4.52 (BERT) and 2.88 (DeBERTa) compared to augmenting the in-domain data. Augmenting both corpora simultaneously (“Synth(both)”) recovers some of this loss but still under-performs in-domain-only augmentation, suggesting that synthetic SMARTSpan notes amplify out-of-domain patterns that compete with the target distribution. The replacement ablation, which swaps 57 SMARTSpan notes for in-domain examples to control for training set size, performs comparably to the unaugmented combined condition, confirming that the improvement from augmentation is not merely a dataset-size effect.

Factorized generation also improves in-domain performance on SMARTSpan.

When both training and evaluation are conducted within SMARTSpan (Table 6, bottom), adding synthetic data generated from the same corpus lifts EM F1 from 90.21 to 91.44 for BERT. For DeBERTa, performance remains near saturation, with EM F1 at 97.08 vs. 97.56 without augmentation. This demonstrates that the pipeline provides value beyond cross-domain adaptation: it also enriches the training in better-resourced, single-domain settings.

Synthetic notes are difficult for humans to distinguish from real ones.

Across 50 trials, evaluators correctly identified the real note with an overall accuracy of 72%. Notably, the evaluators showed little to no preference in terms of reported quality between genuine notes (average quality score: 4.08/5) and synthetic quality (average quality score:

4.06/5). While above chance, this detection rate was reported as largely attributable to a small set of surface-level artifacts rather than to global stylistic differences. In post-evaluation debriefing, both evaluators independently reported relying on the same telltale cues to identify *real* notes: references to named clinics, general practitioners, or community resource centers specific to the local healthcare system; mention of exact appointment dates; the presence of spelling errors and locally conventional abbreviations that the language model did not reproduce; frequent use of clinical shorthand such as *pt*, *hx*, and abbreviated day names; and idiosyncratic formatting habits of individual health coaches. Conversely, the use of em dashes was noted as a recurring indicator of synthetic origin. Notably, evaluators reported that when these surface markers were absent, the two notes were perceptually indistinguishable, suggesting that the factorized pipeline captures the higher-level structure, tone, and goal content of real notes with high fidelity.

4. Discussion

The advantage of factorized generation stems from an inductive bias that mirrors how notes are actually produced: by sampling structure, goal content, and persona independently, the pipeline generates plausible configurations never observed in the seed corpus, a combinatorial diversity that lexical methods like paraphrasing and backtranslation cannot achieve. The disproportionate PM F1 gains suggest this diversity is especially beneficial at span boundaries, where the surrounding context varies more widely across synthetic notes than across paraphrases of the same original.

Critically, the ablation results show that the pipeline acts as a distribution-faithful amplifier rather than a general-purpose data multiplier: augmenting the in-domain seed set enriches the target distribution, whereas augmenting *SMARTSpan* amplifies competing out-of-domain patterns. Practitioners should therefore direct augmentation exclusively at the corpus whose distribution they wish to reinforce, even when out-of-domain data is available for joint training. The human evaluation further confirms that the remaining realism gap is largely cosmetic, driven by missing local abbreviations, absence of spelling errors, and formatting artifacts rather than structural or content flaws, and could likely be narrowed by injecting realistic surface noise during post-processing.

5. Conclusion

We presented a factorized synthetic data generation pipeline that decomposes health coaching note variation into structure, goal content, and per-

sona axes, extracts empirical priors from a small in-domain corpus, and samples from them to produce diverse, label-validated training examples. In low-resource adaptation experiments the pipeline consistently outperformed rephrasing and backtranslation baselines, and ablation analysis showed that targeting augmentation at in-domain data is critical, augmenting out-of-domain data amplifies the wrong distribution. A human evaluation confirmed that synthetic notes are structurally and tonally faithful, with detection driven by surface artifacts rather than content or organizational flaws.

Future work will explore relaxing the cycle-consistency threshold with confidence-weighted training to retain a broader set of synthetic examples, extending the pipeline to multilingual coaching programs, and applying the factorization principle to other forms of structured clinical documentation.

6. Limitations

Several limitations should be noted. First, the pipeline has been evaluated on a single downstream task (goal extraction) with two encoder architectures; its benefit for other information extraction objectives or generative models remains untested. Second, both generation and cycle-consistency validation depend on a single proprietary model (Gemini 3.0 Pro), meaning that systematic biases in that model propagate into the synthetic corpus; whether open-source alternatives achieve comparable factorised generation quality remains to be tested. The strict cycle-consistency threshold mitigates label noise but may preferentially accept simpler notes, subtly skewing the training distribution toward shorter or fewer-goal examples. Third, our new corpus originates from a single coaching program with a predominantly Chinese Singaporean demographic; generalizability to programs with different populations, languages, or documentation conventions has not been established. Finally, the 512-token input limit approaches the mean length of our notes (469 words), and longer notes may lose information at the tail, a constraint that would be alleviated by long-context encoder models.

7. Ethics Statement

The randomized controlled trial from which the health coaching session transcripts and notes were derived received ethics approval from the National Healthcare Group Domain Specific Review Board, Singapore (no. 2023/00438). All participants provided written informed consent prior to enrollment.

8. Acknowledgments

This work was supported by CARDiovascular DiseaseE National Collaborative Enterprise (CADCENCE) National Clinical Translational Program (MOH-001277-01).

9. Bibliographical References

- Zeynab Bahrami, Atena Heidari, and Jacquelyn Cranney. 2022. Applying smart goal intervention leads to greater goal attainment, need satisfaction and positive affect. *International Journal of Mental Health Promotion*, 24(6).
- Iva Bojic, Qi Chwen Ong, Stephanie Hilary Xinyi Ma, Lin Ai, Zheng Liu, Ziwei Gong, Julia Hirschberg, Andy Hau Yan Ho, and Andy WH Khong. 2025. SMARTMiner: Extracting and evaluating SMART goals from low-resource health coaching notes. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16288–16305.
- Julia Bowman, Lise Mogensen, Elisabeth Marsland, and Natasha Lannin. 2015. The development, content validity and inter-rater reliability of the smart-goal evaluation method: A standardised method for evaluating clinical goals. *Australian occupational therapy journal*, 62(6):420–427.
- 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 ACL: human language technologies, volume 1*, pages 4171–4186.
- George T Doran. 1981. There's a smart way to write managements's goals and objectives. *Management review*, 70(11).
- Susan A Flocke and Kurt C Stange. 2004. Direct observation and patient recall of health behavior advice. *Preventive medicine*, 38(3):343–349.
- Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 2020. Human-human health coaching via text messages: Corpus, annotation, and analysis. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 246–256.
- Itika Gupta, Barbara Di Eugenio, Brian D Ziebart, Bing Liu, Ben S Gerber, and Lisa K Sharp. 2021. Summarizing behavioral change goals from sms exchanges to support health coaches. In *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue*, pages 276–289.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Danielle M Hessler, Lawrence Fisher, Vicky Bowyer, L Miriam Dickinson, Bonnie T Jortberg, Bethany Kwan, Douglas H Fernald, Matt Simpson, and W Perry Dickinson. 2019. Self-management support for chronic disease in primary care: frequency of patient self-management problems and patient reported priorities, and alignment with ultimate behavior goal selection. *BMC family practice*, 20(1):120.
- Zixian Huang, Jiaying Zhou, Chenxu Niu, and Gong Cheng. 2023. [Spans, not tokens: A span-centric model for multi-span reading comprehension](#). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, pages 874–884.
- Kirsi Kivelä, Satu Elo, Helvi Kyngäs, and Maria Kääriäinen. 2014. The effects of health coaching on adult patients with chronic diseases: a systematic review. *Patient education and counseling*, 97(2):147–157.
- Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE.
- Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftexhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.
- Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022. Multispanqa: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 1250–1260.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Croía Loughnane, Justin Laiti, Róisín O'Donovan, and Pádraic J Dunne. 2025. Systematic review exploring human, ai, and hybrid health coaching in digital health interventions: trends, engagement, and lifestyle outcomes. *Frontiers in Digital Health*, 7:1536416.
- Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual

- learning for dialogue-act classification in counselling conversations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 735–745.
- National Library of Medicine. 2024. Medical subject headings (mesh) descriptor: Sociodemographic factors. Accessed: 2026-02-19.
- Jeanette M Olsen and Bonnie J Nesbitt. 2010. Health coaching to improve healthy lifestyle behaviors: an integrative review. *American journal of health promotion*, 25(1):e1–e12.
- Zhiyang Qi, Takumasa Kaneko, Keiko Takamizo, Mariko Ukiyo, and Michimasa Inaba. 2025. Kokoro-chat: A japanese psychological counseling dialogue dataset collected via role-playing by trained counselors. *arXiv preprint arXiv:2506.01357*.
- Statista. 2025. [Most common languages used on the internet as of january 2025, by share of web-sites](#). Accessed: 2026-02-19.
- Laurens van der Maaten and Geoffrey Hinton. 2008. [Visualizing data using t-SNE](#). *Journal of Machine Learning Research*, 9(86):2579–2605.
- Anne M Wallace, Matthew T Bogard, and Susan M Zbikowski. 2018. Intrapersonal variation in goal setting and achievement in health coaching: cross-sectional retrospective analysis. *Journal of Medical Internet Research*, 20(1):e32.
- Nicole D White, Vicki Bautista, Thomas Lenz, and Amy Cosimano. 2020. Using the smart-est goals in lifestyle medicine prescription. *American journal of lifestyle medicine*, 14(3):271–273.
- Ruth Q Wolever, Leigh Ann Simmons, Gary A Sforzo, Diana Dill, Miranda Kaye, Elizabeth M Bechard, Mary Elaine Southard, Mary Kennedy, Justine Vosloo, and Nancy Yang. 2013. A systematic review of the literature on health and wellness coaching: defining a key behavioral intervention in healthcare. *Global advances in health and medicine*, 2(4):38–57.
- World Health Organization. 2019. Self-care interventions for health: global evidence and recommendations. Technical report. Accessed: 2026-02-19.
- Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. MentalChat16K: A benchmark dataset for conversational mental health assistance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5367–5378.
- Yue Zhou, Barbara Di Eugenio, Brian Ziebart, Lisa Sharp, Bing Liu, and Nikolaos Agadakos. 2024. Modeling low-resource health coaching dialogues via neuro-symbolic goal summarization and text-units-text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11498–11509.
- Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

A. Example Health Coaching Notes

Figures 4 and 5 present two real and two synthetically generated health coaching notes, respectively. Goal spans are highlighted in green. The synthetic notes were produced by the factorised generation pipeline, sampling structure, goal attributes, and persona independently from the empirical priors extracted during corpus analysis (Section 2.2.1).

B. Corpus Analysis Prompts and Schemas

All corpus analysis calls (Section 2.2.1) use Gemini 3.0 Pro with temperature 0 and Pydantic-enforced structured generation. Figures 6–8 reproduce the system prompt, user prompt template, and output schema for each analysis axis.

29 Dec 2025 (1st session). Current goals: Exercise: Jogging 30 minutes every Saturday around 5pm at the park near his house. 6/10 confident. Medication: To take medication regularly. Morning and night. 9/10 confident. Wellness Vision: He hopes to maintain his current levels of health, in terms of blood pressure, cholesterol and blood glucose. He does not want to develop complications or stroke due to the previous mention conditions. He is afraid that he will suffer and have his quality of life affected. 3 month goal To lose weight. Notes: Deals with company operations. Works normal working hours. Married with 2 children. One of his son is attending health coaching sessions. He is doubtful about the effectiveness of health coaching as he does not believe one would have the luxury of time to make lifestyle changes. He has high blood pressure and cholesterol. There are currently within target as he is compliant with medication. He had previously attempted lifestyle changes like diet and exercise. While they did reduce his cholesterol levels, it was not enough. He therefore agreed to start on medication. His original concerns were that once a person starts medication it is for life. He understands the importance of keeping his blood cholesterol and pressure in check. He does not want to get a heart attack or stroke. Getting a serious illness is fine as long as he passes away straight. He expressed concern about not being able to provide for his family and the suffering one has to go through after getting a stroke. Some changes to his lifestyles were eating only one meal a day and exercising on weekends. As he is now on medication, he admitted that he has not been that consistent with his healthy habits as medication has been keeping his health in check. He was previously maintaining the healthier lifestyle for one to 2 years. While the results were not immediate, he was able to pursue on, and the improvements motivated him to continue being consistent. Improvements were seen in blood test results and reduce clothing sizes. He managed to lose 4kg. He was also previously a smoker and stopped smoking when he developed high blood pressure. What was helpful in today's session was the realization about the accomplishments he had made in the past regarding his lifestyles. That it can be accomplished and how much he has changed.

Session 3 (08/07/2025): (1) Exercise: Continue with muscle strengthening exercises (leg lifts and plants) daily (7/10), (2) Diet: Eat 2 meals a day for 5 days at home (7/10), (3) Diet: No fried chicken (KFC/Tenderbest) for the month, even if there is a good offer (10/10). Preferred name: Sok Leng Wellness vision: Free of new chronic illness, managing weight, being independent, and maintaining physical health to enjoy the things I like. 3m goal: 77kg - 74kg Highlight of the month: Her pipe burst at home, she took 2 weeks to resolve it because she needed to source for plumbers who can fix it at a reasonable price. Due to that, it affected her health goal of cooking at home. Another highlight was a temporary role as an invigilator/usher for an international math olympiad where she was tasked to care for children in Pri 1 - 3. She moved around a lot during those times and felt that her activity levels increased. GR 1: Muscle strengthening (60%) completion. She was glad that she started the exercises and that had helped with the muscle aches. She is motivated by her parent's experience (where they require support moving around), and would like to work on preventive measures to support mobility. The pipe incident at home did not affect her ability to exercise as she believed that it was a small matter and did not allow it to affect her. GR2: Diet (60%) completion. Her pipe burst incident affected her goal of making home cooked food but she was able to take healthier, home-cooked meals when she was visiting her parents. She shared that there were days where she had KFC fried chicken for her meals because of a bundle deal that was hard to resist. She also shared her appreciation for a deep-fried chinese bun that was sold a few MRT stops from her place, which she buys when there is a sale. Other: In the later part of July, she will be travelling to Shenzhen (till 26th), followed by a university trip to Guizhou (27th onwards). It was a learning trip that she was keen to experience. Action plan: Recommend food videos. Continue to work around adjusting diets - making some progress on this. In discussion, she shared that she feel less aches and pains after trips as compared to the past - this may be due to the terrain, but if she raises it again, perhaps focus on the enabling factors or what was different.

Figure 4: Two real health coaching notes from our corpus. Goal spans are highlighted in green. Top: 2 goals. Bottom: 3 goals.

Session 3 (14/08/2025). Patient is a young adult male, financially stable but finds his current work challenging due to constantly chasing payments and meeting high demands. He is reflective about his health and life direction. Known case of familial hypercholesterolemia, currently managing with medication and herbal supplements. Lives in an area with good access to food stalls and amenities, and has running gear available at home. Wellness Vision: To feel energetic and in control of my body, reducing reliance on long-term medication. 3-month goals: 1. Improve lipid profile naturally 2. Establish a consistent running routine 3. Clean up diet significantly. We discussed his medication adherence; he is compliant but expressed a strong desire to eventually stop taking statins by proving he can manage his cholesterol through lifestyle changes alone. He understands this requires significant dietary discipline. Barriers: Work stress often leads to emotional eating or skipping workouts. He notes that when he is chasing payments, his anxiety spikes, making him reach for comfort foods. Technical issues: He had trouble syncing his wearable device earlier this week but resolved it; energy levels have been fluctuating with work stress. He feels drained by 6 PM most days. Barriers regarding exercise were discussed: previously, rain and fatigue were issues, but he has access to a treadmill now. Goals set for this period: Diet: To consume No chips and no ice cream; substitute with fruits and tea (Confidence 8/10). Diet: No fried chicken (KFC/Tenderbest) 3 per week. Exercise: To do a run on the treadmill / pool walking 1 per day (Confidence 6/10). Regarding diet, I noted that he tends to eat out at hawker centers often. He has been trying to manage his carbohydrate intake by asking for less rice, but admits the gravy makes it hard to gauge calories.

15 Oct 2025 (Session 4). Goals: Physical Activity: Meet 5k steps by jogging and brisk walking. He plans to do this 2 times per day for a total duration of 8 hours during his night shifts. 10/10 confident. This is a strategy to combat the fatigue he feels during the long remote shifts. Diet: Abstain from fried food (e.g. fried chicken, potato chips) in the morning for the next 3 months. 8/10 confident. He noted that after his night shift ends, he often craves greasy breakfast foods with his wife, but wants to break this cycle. Generative moment: The client reflected on his recent interactions with his elderly father, who has limited mobility. Seeing his father struggle to walk has been a wake-up call for him. He realized that his sedentary job and previous lack of activity were leading him down a similar path. He expressed a strong desire to not be reliant on medication like his father is. This emotional connection to his family's health history seems to be a major driver for his current motivation. Longterm goals review: His wellness vision is to maintain a healthy weight and good energy levels so he can enjoy cycling on the weekends without feeling exhausted. He wants to reduce his blood pressure medication dosage within the next year. Patient information: Client will continue to navigate the challenges of working night shifts from home. He will likely face peer pressure from colleagues who suggest ordering fast food during virtual meetings, but he intends to have meal prep ready. He will also be managing the care of his elderly parents on his days off. Other points: He originally joined the program because a nurse at his annual screening flagged his rising blood pressure and weight. He admitted that working from home made him complacent, often sitting for 10-12 hours a day. He is currently experiencing significant fatigue, which he attributes to the shift work and poor diet. App usage review: He will use the app to track his steps during his shift. He will also watch the educational videos on managing shift work sleep patterns to help him get better rest.

Figure 5: Two synthetically generated health coaching notes. Goal spans are highlighted in green. Top: 3 goals. Bottom: 2 goals. Both notes were produced by the factorised generation pipeline.

Goal Attribute Enrichment

PROMPT

System:

You are an expert at analyzing health coaching goals. Your task is to extract detailed attributes from SMART goals (Specific, Measurable, Achievable, Relevant, Time-bound).

Extract attributes precisely from the given goal text. If information is not present, return null for that field. Be accurate and don't infer information that isn't explicitly stated.

User template (formatted per goal span, with full note as context):

Goal to analyze: "{goal_text}"

Full context from coaching note:
{note_context}

Extract the following attributes from this goal:

1. goal_type: Type of health behavior goal (physical_activity, dietary_behavior, medication_adherence, substance_use, self_monitoring)
2. goal_activity: The main activity or behavior (e.g., 'jogging', 'cooking', 'taking medication')
3. location: Where the activity takes place (e.g., 'at home', 'gym', 'park')
4. time: Time of day or specific time (e.g., 'morning', 'evening', 'after work')
5. frequency: How often (value, unit, and full text)
6. duration: How long per session (value, unit, and full text)
7. timeframe: When/for how long (type and full text)
8. conditionality: Any conditional aspects
9. confidence: How confidence is expressed (type, value if numeric, and full text)
10. zero_goal_reason: If "Goal to analyze" is empty (""), select one of:
no_show | not_discussed_declined | not_discussed_review_only | other

Return null for any attribute not present in the text, EXCEPT for zero_goal_reason which must always have a value.

OUTPUT SCHEMA

Field	Type	Description
goal_type	Optional[Literal]	<i>physical_activity dietary_behavior medication_adherence substance_use self_monitoring</i>
goal_activity	Optional[str]	Main activity or behavior
location	Optional[str]	Where the activity takes place
time	Optional[str]	Time of day or specific time
frequency_value	Optional[int]	Numeric frequency value
frequency_unit	Optional[Literal]	<i>week day month year null other</i>
frequency_text	Optional[str]	Full frequency text as written
duration_value	Optional[int]	Numeric duration value
duration_unit	Optional[Literal]	<i>hours minutes days weeks months null other</i>
duration_text	Optional[str]	Full duration text as written
timeframe_type	Optional[Literal]	<i>specific_day date_range ongoing other</i>
timeframe_text	Optional[str]	Full timeframe text
conditionality	Optional[str]	Conditional aspects
confidence_type	Optional[Literal]	<i>numeric_percent numeric_scale none</i>
confidence_value	Optional[float]	Numeric confidence value
confidence_text	Optional[str]	Full confidence text
zero_goal_reason	Optional[Literal]	<i>no_show not_discussed_declined not_discussed_review_only other</i>

Goal types follow the taxonomy of Hessler et al. (2019). Schema enforced via Pydantic structured generation.

Figure 6: Prompt and output schema for goal attribute enrichment (Section 2.2.1). Goal types follow the taxonomy of Hessler et al. (2019).

Structural Analysis

PROMPT

System:

You are an expert at analyzing the structure and formatting of health coaching notes. Your task is to identify the structural patterns used in coaching notes.

Extract structural attributes precisely from the given note text. Be accurate and classify based on what is actually present in the text.

IMPORTANT: Be thorough and granular. Most coaching notes contain many distinct sections (typically 4-10). Each time the note shifts to a different topic, that is a separate section. Do NOT lump multiple topics into a single section. For example, if a note discusses medication, then food, then exercise as separate paragraphs or blocks, each of those is its own section – even if they lack explicit headers.

User template (formatted per note):

Coaching note to analyze:
{note_text}

Analyze the structural pattern of this coaching note. Identify EVERY distinct section – a new section begins whenever the note shifts to a different topic or purpose, whether or not there is an explicit header.

For each section, extract:

- * "type": weekly_goals_review, longterm_goals_review, app_usage_review, journalling_review, goal_setting, barriers_discussion, patient_information, generative_moment, or other_points
- * "format": numbered_list, embedded_narrative, or unordered_list
- * "voice": third_person, first_person, coach_we, passive, or mixed
- * "tense_pattern": mixed_past_future, primarily_future, or primarily_past
- * "section_summary": Optional short summary (only for app_usage_review, journalling_review, longterm_goals_review, weekly_goals_review, other_points)

Important guidelines:

- Be GRANULAR: if the note discusses medication adherence, then diet, then exercise as separate blocks, those are THREE separate sections – not one combined section.
- The same section type CAN appear multiple times.
- Typical coaching notes have 4-10 sections. If you find only 1-2 sections in a long note, re-read it more carefully.
- Return empty list ONLY if the note has no clear sections at all.

OUTPUT SCHEMA (LIST[SECTION])

Field	Type	Description
type	Literal	Section type from checklist <i>weekly_goals_review longterm_goals_review app_usage_review journalling_review goal_setting barriers_discussion patient_information generative_moment other_points</i>
format	Literal	<i>numbered_list embedded_narrative unordered_list</i>
voice	Literal	<i>third_person first_person coach_we passive mixed</i>
tense_pattern	Literal	<i>mixed_past_future primarily_future primarily_past</i>
section_summary	Optional[str]	Short summary (only for review/other_points types; null otherwise)

Section types derived from the health coaching session checklist used in clinical practice.

Figure 7: Prompt and output schema for structural analysis (Section 2.2.1). Section types are derived from the health coaching session checklist used in clinical practice.

Persona Extraction

PROMPT

System:

You are an expert at analyzing health coaching notes to identify persona elements. Your task is to extract individual persona characteristics present in the note.

Extract persona elements based on what is stated or reasonably inferred from the note. Be thoughtful in your inferences but avoid being overly creative – stick to what the text supports.

User template (formatted per note):

Coaching note to analyze:
{note_text}

Extract persona elements present in this note across the following categories:

1. Social demographics:
 - age_group: young_adult (18-35), middle_aged (36-60), senior (60+), or unknown
 - occupation: Current occupation or employment status (free-form)
 - education_level: Education level (free-form)
 - gender: Gender identity (free-form)
 - ethnicity: Ethnicity or racial background (free-form)
 - income_level: Income level or socioeconomic status (free-form)
 - marital_status: Marital status (free-form)
2. Social context: Social relationships, social factors, and social barriers that influence health behaviors (e.g., lives_with_partner, social_eater, family_dinners, social_triggers)
3. Economic context: Economic and financial factors, employment-related circumstances, and economic barriers (e.g., cost_concerns, financial_stress, late_shifts, work_demands)
4. Physical environment: Physical environment, living situation, and physical barriers (e.g., lives_in_urban_area, access_to_gym, limited_kitchen_space, travel_disruptions)
5. Individual characteristics and behaviours: Behavioral patterns, health literacy, psychological state, and individual-level barriers (e.g., all_or_nothing_thinker, routine_dependent, tech_comfortable, time_constraints)

Guidelines:

- Use snake_case naming for all extracted elements
- Return empty lists for categories where no elements are present
- Return null for demographics not available
- When barriers are mentioned, categorize them into the appropriate context category based on their nature

OUTPUT SCHEMA

Field	Type	Description
social_demographics (SocialDemographics)		
age_group	Optional[Literal]	young_adult middle_aged senior unknown
occupation	Optional[str]	Occupation or employment status
education_level	Optional[str]	Education level
gender	Optional[str]	Gender identity
ethnicity	Optional[str]	Ethnicity or racial background
income_level	Optional[str]	Income level or socioeconomic status
marital_status	Optional[str]	Marital status
Context and behavioural factors (all List[str], snake_case)		
social_context	List[str]	Social relationships and social barriers influencing health behaviours
economic_context	List[str]	Economic/financial factors and employment-related barriers
physical_environment	List[str]	Physical environment, living situation, and environmental barriers
individual_characteristics_and_behaviours	List[str]	Behavioural patterns, health literacy, psychological state, individual barriers

Demographic categories informed by MeSH Sociodemographic Factors and WHO self-care frameworks.

Figure 8: Prompt and output schema for persona extraction (Section 2.2.1). Demographic categories informed by MeSH Sociodemographic Factors and WHO self-care frameworks.

Scalable Generation of Adult-Oriented Therapeutic Reading Texts for Russian Aphasia Rehabilitation

Anastasia Kolmogorova, Anastasia Margolina, Alina Telnova and Igor Ilchenko

National Research University Higher School of Economics, Russia

akolmogorova@hse.ru, avmargolina@edu.hse.ru, aatelnova@edu.hse.ru, igvladilchenko@gmail.com

Abstract

Texts are widely used in aphasia rehabilitation to support the recovery of comprehension and narrative planning. In routine practice, clinical impact depends strongly on patient motivation and on the availability of age-appropriate reading materials: adults are often offered child-oriented texts, which can be perceived as demeaning and may reduce engagement. We present a controllable generation pipeline for building a repository of Russian therapeutic reading texts for adult aphasia therapy. An anonymized repository with code and data is available at <https://github.com/z00logist/aphasia-exercises-generation>. The pipeline conditions each story on an explicit semantic triplet (12 topics $\times$ 10 archetypes $\times$ 11 objects) and enforces three clinically motivated complexity regimes (Basic/Intermediate/Advanced). Using batched prompting, we generate 1,296 unique stories. We evaluate the corpus with classical linguistic metrics, sentiment analysis, and a LLM-as-a-judge protocol (18 binary criteria); on a stratified sample of 198 stories, overall rubric compliance is 80.0%. Surface metrics show a monotonic increase in lexical and syntactic complexity across regimes, and comparisons against a clinical anchor set of 10 therapist-authored texts reveal that while generated texts match ground truth readability at the Basic level, standard clinical texts align closer to the Intermediate generated level in length and complexity. Judge-based analysis indicates near-perfect adherence to high-level narrative constraints but persistent limitations in fine-grained phonotactic control, which we manually confirm, motivating hybrid neuro-symbolic enforcement.

Keywords: aphasia therapy, controlled text generation, Russian, readability, LLM-as-a-judge, evaluation

1. Introduction

Texts are widely used by speech-language pathologists (SLPs) in speech therapy interventions for both children and adults presenting with various speech disorders or language impairments. Aphasia is one such systemic language disorder resulting from brain damage caused by traumatic injury, stroke, or surgical intervention (e.g., tumor resection) (Luria, 2008).

In speech therapy, texts serve as a tool for restoring different functional skills: for supporting the formation and execution of simple motor programs (e.g., during reading-while-listening tasks), for restoring connected text comprehension (e.g., tasks focused on understanding content), for vocabulary expansion, and for rehabilitating utterance planning skills (e.g., retelling) (Luria, 2008).

However, clinical practice indicates that a patient's motivation to read plays a crucial role in the effectiveness of text-based interventions in speech therapy (Webster et al., 2023). Motivation is shaped by several factors: whether the subject matter is engaging for the patient; whether linguistic complexity matches the patient's current functional level; and whether the text leaves a positive emotional impression after reading (Webster et al., 2023). Nevertheless, such texts are not read for entertainment, but rather as a means to achieve specific rehabilitation goals set by the SLP. The SLP, in turn, uses these texts to assess progress and plan subsequent therapy. We there-

fore treat these materials as texts for specific purposes, namely speech and language recovery.

Selecting thematically relevant texts for patients and having them expertly simplified is a task that exceeds the capacity of a practicing speech therapist in routine clinical settings (Hoover et al., 2023). While engaging reading programs show promise (Cocks et al., 2013), manual customization to adult interests and impairment-specific linguistic levels remains difficult to scale. We address this gap by presenting a sustainable computational pipeline for generating texts with specified topics, genre structures, and linguistic complexity, thereby creating an expandable repository of such materials.

2. Therapeutic Texts for Aphasia: Topic, Structure, and Complexity

To understand the specific features of texts used in aphasia therapy as a genre, we compiled a corpus of 127 texts. These were drawn from textbooks and methodological guides for speech-language therapists working with aphasia, as well as from texts employed by practicing therapists, selected by them to meet the needs of individual patients.

Further analysis revealed that texts designed for patient work predominantly feature child-oriented themes: parent-child relationships, children's games, and grandparents with grandchildren. However, in texts individually selected by speech therapists, the topics are already more closely

aligned with patients' age and life experience: cars, hobbies, gardening. In expert interviews, therapists noted that patients feel offended when offered child-oriented reading materials. We compiled a list of 12 topics designed to represent ecological domains of adult life: Domestic Activities, Modern Technology, Gardening, Hobbies, Cooking, Travel & Transport, Shopping & Finances, Health, Socializing, Pets, Work & Profession, and Daily Routine. This selection prioritizes subjects that facilitate functional communication and social engagement, purposely excluding child-oriented themes common in traditional therapy.

In terms of composition, texts for speech therapy are structured with remarkable clarity: exposition, climax, resolution.

Interestingly, there is a certain tonal movement within the text: it may begin neutrally, then acquire a slight negative shading, but the final part is always positive and contains a takeaway — a kind of lesson or moral arising from the situation described. The text prompts reflection, observation, or gentle irony — not didacticism, but a soft philosophical undertone. Meaning is conveyed through action or image rather than direct moralizing, allowing the reader to infer that "this is about us":

Example translated to English. A gardener had two sons. The gardener owned a large vineyard. The sons did not want to work in the vineyard. One day, the father called his sons and said, "Children, go into the vineyard and look for something hidden there." The sons thought a treasure was buried in the vineyard. They went and began digging the ground. They dug up the entire vineyard, but found no treasure. However, the soil became loose and fertile. The vineyard produced abundant grapes. The sons sold the grapes and became rich.

Each text corresponds to one of the following genres: a real-life incident, a parable, a fairy tale, or a didactic story. There is a conflict or contrast (misunderstanding, change, confrontation). The hero loses something, discovers something, or undergoes a change (internal or external). The conflict is neither dramatic nor tragic. The number of characters in the texts is small — typically one or two main protagonists. To systematize them for generation purposes, we drew on the principles of the morphology of the folktale as developed by the Russian formalist Vladimir Propp (Propp, 1968). He argued that despite the immense variety of fairy-tale plots, seven generalized character types can be identified. An analysis of existing therapy texts allowed us to distinguish ten such character archetypes. They reflect the core gender and age oppositions that structure society, as

well as two basic social roles from the "inner circle" (neighbor and friend): Old Man, Old Woman, Man, Woman, Boy, Girl, Friend, Neighbor, Young Man, Young Woman. In addition, the characters interact with some objects from the everyday environment. However, the object with which a character interacts is not merely named in the text — it is assigned a certain attribute. To systematize the attribute and its bearer for text generation, we employed a minimal predicative unit: the "noun + adjective" pair. The choice of the predicative word was likewise not arbitrary: according to the theory of argumentation within language developed by Jean-Claude Anscombre and Oswald Ducrot (Anscombre and Ducrot, 1983), every linguistic community possesses prototypical representations of the properties that objects and people can have. Thus, prototypical work is hard, and a prototypical relative is close. Drawing on our own introspection as native speakers and on association measures between collocates in the Russian National Corpus, we formulated eleven such "object — prototypical attribute" pairs: red ball, big book, old key, wooden box, heavy stone, important letter, old photo, yellow flower, blue cup, green apple, and new hat.

Regarding the linguistic complexity of texts for speech therapy, its parameters correspond to the severity of the patient's language impairment. However, in clinical practice, speech therapists adjust these parameters intuitively, based on their own expertise. In our previous work, we explicitly formulated three types of linguistic complexity requirements tailored to patients with mild, mild-to-moderate, and moderate impairment severity (Author and Author, 2024). In the following text, we will refer to these levels as Basic, Intermediate and Advanced, respectively. These requirements were developed through iterative collaboration with expert speech therapists and were clinically validated. When discussing the generation results, we will draw, in part, on these text requirements.

Thus, the analysis of texts already used in therapy, combined with input from speech therapists, allowed us to formulate the following aspects and characteristics of this text type, which will subsequently inform both the generation and validation of the output. Table 1 summarizes the target properties used as design constraints.

3. Methods

The input space is modelled as a semantic triplet (c, a, o) , where c is a thematic category (12 types), a is a character archetype (10 types), and o is a key object (11 types). This combinatorial space (1,320 unique triplets) ensures thematic diversity

#	Aspect	Description
1	Structure	Clear three-part structure: exposition, climax, resolution
2	Character	10 archetypes reflecting core social roles
3	Key Object	11 object-attribute pairs for interaction
4	Conflict	Conflict or contrast (misunderstanding, change, confrontation)
5	Depth	Philosophical undertone / gentle reflection
6	Emotion	Socially approved emotions dominate
7	Genre	Real-life incident, parable, fairy tale, or didactic story
8	Language	3 regimes: Basic, Intermediate, Advanced
9	Topic	12 adult-oriented topics

Table 1: Target aspects and characteristics of therapeutic texts for aphasia.

while avoiding the infantilized narratives typical of accessible reading materials.

We operationalize clinical progression via three complexity regimes defined by constraint bundles. Basic texts are strictly limited to 5–7 simple past-tense SVO sentences with no dialogue. Intermediate narratives introduce limited subordination within a 2–3 paragraph structure, while Advanced texts permit complex syntax, descriptive modifiers, and abstract themes. This approach ensures clinical interpretability across diverse topics.

To validate these regimes, we test two hypotheses and one research question: (H1) complexity regimes form statistically distinct linguistic distributions; (H2) Basic texts align quantitatively with a clinical anchor; and (Q3) high-level narrative constraints require LLM-based assessment. Following (Alva-Manchego et al., 2020), we assess whether the regimes form statistically separable distributions across readability and syntactic indicators. To capture tonal alignment, we evaluated sentiment distributions using the `cointegrated/rubert-tiny-sentiment-balanced` model. Since surface statistics cannot capture narrative structure or tonal appropriateness, we complement them with a deterministic LLM-as-a-judge pipeline using binary criteria, interpreting results with awareness of potential model biases (Zheng et al., 2023).

4. Text Generation Pipeline

We generate texts via batched prompting using Gemini 1.5 Flash (Reid et al., 2024) through the Google Generative AI API (Figure 1). We selected this model because it matched the operational requirements of the generation stage:

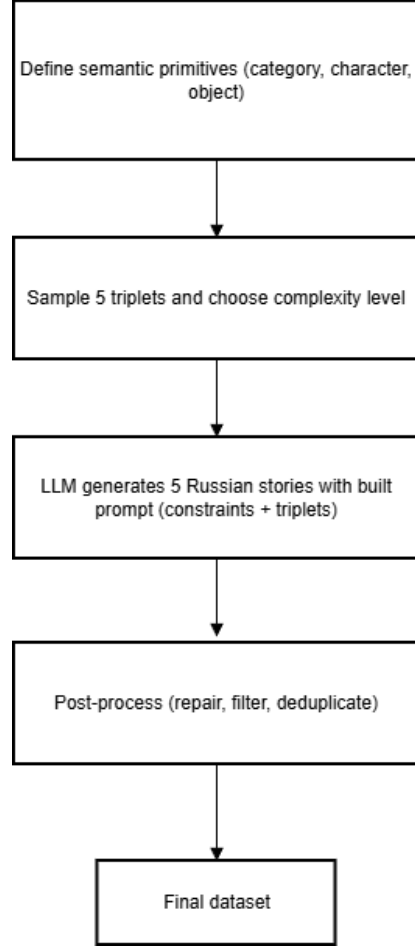


Figure 1: Synthetic dataset creation pipeline producing 1,296 unique stories.

stable schema-following behavior, efficient API-based batched inference, and sufficiently reliable instruction-following under Russian prompt conditions. Since Google does not publicly disclose the parameter count of Gemini 1.5 Flash, we avoid making capacity claims based on model size and motivate its use empirically through pilot-stage generation behavior instead. For each batch, we fix one complexity regime, sample five unused triplets, and prompt the model to produce one story per triplet under regime-specific constraints. We use deterministic or low-variance decoding and enforce a fixed output schema in the prompt.

Large-scale generation produces recurrent formatting failures. We observe CSV breaks caused by commas inside the generated text field; we repair these by merging spillover columns back into the text column until the expected schema is restored. We remove empty outputs and deduplicate near-identical stories.

After post-processing, the corpus contains 1,296 unique stories: 31.0% Basic, 33.7% Intermediate, and 35.3% Advanced.

Furthermore, the analysis revealed a high de-

gree of corpus balance with respect to content parameters as well. All 12 thematic categories are uniformly represented, each accounting for between 8.0% and 8.6% of the total number of texts. A similarly even distribution is observed for the 10 character archetypes and the 11 key objects. This balance ensures that when the corpus is used in therapeutic practice, the risk of systematic bias toward any particular theme, character, or object is minimized, thereby promoting greater variability and potential patient engagement.

5. Evaluation

We benchmark against a “Ground Truth” (GT) set of 10 manually authored therapy texts used in current clinical practice. These texts were selected by practicing therapists based on their everyday utility. They were not pre-selected to match any one generated regime; rather, we use them as a small clinical anchor representing materials currently employed in routine therapeutic work.

5.1. Classical Linguistic Metrics

We evaluated the full synthetic corpus and used the 10 GT texts to contextualize the regimes. We report readability (SMOG [McLaughlin, 1969](#); Russian Flesch Reading Ease [Oborneva, 2006](#)), lexical diversity (MTLD [McCarthy, 2005](#)), and syntactic complexity (mean dependency distance, MDD [Liu, 2008](#)), alongside surface statistics.

As shown in Table 2, the results exhibit a monotonic progression from Basic to Advanced across all indicators, complete with standard deviations to capture within-level variability.

Regarding H2, the metrics reveal an important nuance: the Ground Truth texts do not uniformly align with the Basic generated level. While the GT texts closely match the Basic level in readability (Flesch RE: 68.0 vs. 65.0), they align much more closely with the *Intermediate* generated level in Word Count (92.8), SMOG (13.2), MTLD (123.8), and MDD (2.73). This indicates that the automatically generated “Basic” texts are structurally simpler (shorter and with lower dependency distance) than the therapist-authored clinical anchor. We therefore interpret the Basic regime as a lower-complexity option that may be more accessible for patients with more severe aphasia, rather than as a direct approximation of standard therapist-authored materials.

Additionally, sentiment analysis highlighted a distinct tonal shift. The human-authored GT texts feature notable negative dramatic tension (Negative Sentiment: 0.500 ± 0.33 , Compound: -0.35), typical of fables that establish a problem before a resolution. The generated texts, however,

Metric	Basic	Interm.	Adv.	GT
<i>Surface Indicators</i>				
Word Count	23.0 $\pm$ 3.1	59.2 $\pm$ 9.2	111.0 $\pm$ 31.5	92.8 $\pm$ 35.6
ASL (words/sent)	3.4 $\pm$ 0.5	10.3 $\pm$ 2.4	13.2 $\pm$ 2.7	9.2 $\pm$ 2.8
<i>Readability</i>				
Flesch RE (RU)	65.0 $\pm$ 24.7	52.6 $\pm$ 12.8	49.3 $\pm$ 11.0	68.0 $\pm$ 9.2
SMOG	9.5 $\pm$ 1.7	14.6 $\pm$ 1.7	16.0 $\pm$ 1.7	13.2 $\pm$ 2.0
<i>Lexical & Syntactic</i>				
MTLD	23.4 $\pm$ 17.3	137.1 $\pm$ 66.7	154.2 $\pm$ 42.1	123.8 $\pm$ 56.3
MDD	1.54 $\pm$ 0.15	2.83 $\pm$ 0.26	3.02 $\pm$ 0.24	2.73 $\pm$ 0.31
Subclauses / sent	0.03 $\pm$ 0.09	0.87 $\pm$ 0.38	0.99 $\pm$ 0.39	0.50 $\pm$ 0.28
<i>Morphological</i>				
Past Tense Ratio	0.98 $\pm$ 0.09	0.81 $\pm$ 0.14	0.79 $\pm$ 0.15	0.72 $\pm$ 0.15
Errors / 100w	0.04 $\pm$ 0.43	0.31 $\pm$ 0.88	0.27 $\pm$ 0.54	1.32 $\pm$ 3.60

Table 2: Mean linguistic metrics ($\pm$ SD) by complexity regime and ground-truth (GT) texts.

Metric	H	p
Average Sentence Length (ASL)	936.6	< 0.001
Mean Dependency Distance (MDD)	882.8	< 0.001
SMOG Readability Index	872.8	< 0.001

Table 3: Statistical validation of complexity stratification using the Kruskal–Wallis H test.

lean heavily positive across all levels (Compound scores ranging from +0.22 to +0.30). To confirm that these observed differences reflect statistically distinct distributions rather than incidental variation, we applied the Kruskal–Wallis H test to core metrics ([Kruskal and Wallis, 1952](#)). As shown in Table 3, differences across regimes are highly significant for all tested indicators (all $p < 0.001$), supporting H1.

At higher levels, the model drifts toward a syntactically denser narrative style than the clinically curated GT: Intermediate and Advanced texts show substantially higher subordination rates (0.87 and 0.99 subclauses per sentence) than the GT (0.50). This suggests that, without explicit constraints, the model does not reliably preserve the unusually sparse syntax characteristic of the therapist-authored clinical anchor. This syntactic drift motivates complementary, structure-aware evaluation beyond surface statistics (Q3; see Section 5.2).

5.2. LLM-as-a-Judge

To address Q3, we implemented an automated evaluation pipeline using Gemma-3-4b-it to assess narrative coherence, tonal appropriateness, and fine-grained linguistic constraints ([Gemma Team et al., 2024](#)). We selected Gemma-3-4b-it as a compact open 4B instruction-tuned model for repeated rubric-based evaluation. Gemma 3 supports multilingual processing, including Russian, and provides a 128K context window, making it practical for consistent judge-style assessment across our sampled corpus. We treat these

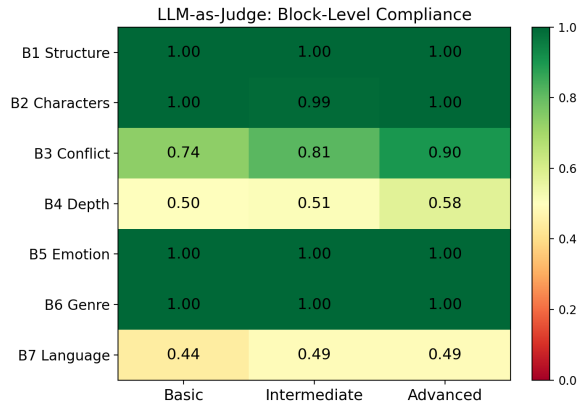


Figure 2: Heatmap of LLM-as-a-Judge compliance scores across seven conceptual blocks. (Note: 'short', 'medium', and 'long' in the figure correspond to the Basic, Intermediate, and Advanced regimes, respectively). Structural and emotional constraints show near-perfect adherence, while Language Constraints (B7) and Semantic Depth (B4) reveal limitations.

judge scores as exploratory and complementary to classical metrics, not as a substitute for expert clinical assessment. We designed a rubric comprising 18 binary criteria organized into seven conceptual blocks (aligning directly with the aspects in Table 1). A random sample of 198 stories (stratified by complexity level, 66 per level) was evaluated using deterministic decoding to ensure reproducibility. The prompt explicitly instructed the LLM-as-a-Judge model to act as a harsh critic, verifying constraints specific to the Russian language (Zheng et al., 2024). Note that some criteria evaluated here (e.g., presence of reflection/irony, strict phonotactics) were not explicitly mandated in the generation prompt; this evaluation serves to identify natural gaps in LLM outputs to inform future neuro-symbolic interventions, rather than solely grading prompt adherence.

The overall compliance rate across the sampled corpus was 80.0%. At the block level (Figure 2), the model demonstrated near-perfect performance on high-level narrative instructions. Compliance scores for Narrative Structure (B1), Emotional Tone (B5), and Genre Appropriateness (B6) were nearly absolute across all levels. Notably, the Conflict & Dynamics block (B3) revealed a complexity-dependent progression: the model struggled to establish meaningful conflict within the tight constraints of Basic texts but successfully developed narrative tension when afforded greater length in Advanced stories.

However, the criteria-level analysis (specifically within B7: Language Constraints, see Figure 3) highlights specific limitations in controlling fine-grained linguistic features. The criterion prohibit-

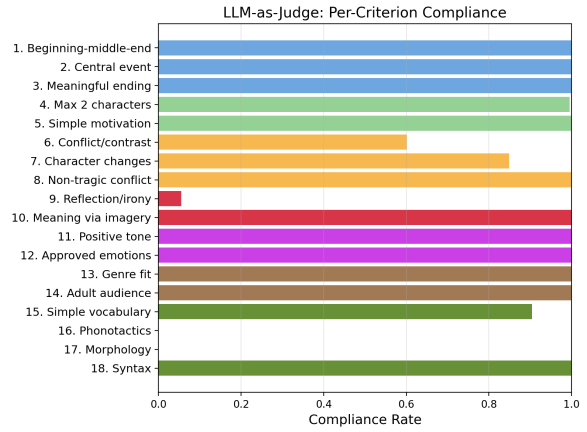


Figure 3: Detailed compliance rates by individual criterion (1–18). The chart illustrates the disparity between high-level narrative success and low-level linguistic control (e.g., Phonotactics).

ing consonant clusters of three or more sounds (Phonotactics) yielded 0% compliance from the judge. This criterion is important because articulating such consonant clusters is problematic for patients with motor aphasia. To validate this result, we manually checked the corpus using a deterministic regex-based script. The manual analysis showed that actual compliance was 18.6% overall (Basic: 52.2%, Inter.: 6.6%, Adv.: 0.4%). Thus, although the generator does struggle with this constraint, the judge substantially underestimates compliance on this character-level task. We interpret this as evidence that LLM-based evaluators are unreliable for phonotactic verification of this kind and should be supplemented with deterministic symbolic checks.

Additionally, a discrepancy emerged in Morphology: while classical metrics confirmed high past-tense usage, the judge's binary scoring penalized even single deviations, resulting in a floor effect. Finally, the low score on Reflection/Irony suggests that while the model generates coherent therapeutic scenarios, it lacks the emergent semantic depth required to produce the philosophical subtext characteristic of human-authored fables without few-shot guidance.

Judge comparison against the clinical anchor.

To better contextualize the judge-based scores, we also applied the same 18-criterion rubric to the 10 therapist-authored ground-truth (GT) texts currently used in clinical practice. As shown in Figure 4, the GT texts achieved an overall compliance rate of 83.3%, compared to 80.0% for the generated sample. This result is important for two reasons. First, it shows that the generated corpus remains close to the clinical anchor at the aggregate level under this rubric. Second, it reveals lim-

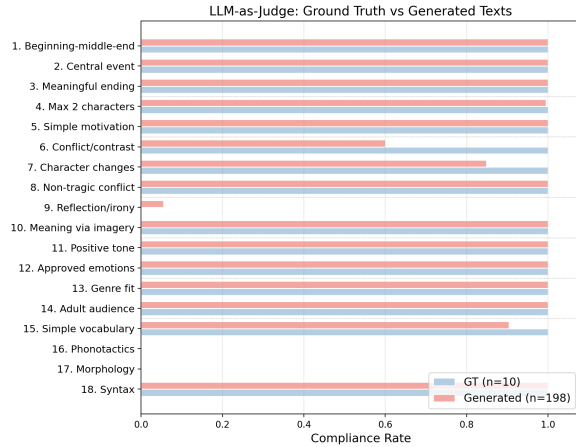


Figure 4: LLM-as-a-Judge comparison between therapist-authored ground-truth (GT) texts and generated texts. GT attains a slightly higher overall compliance rate (83.3% vs. 80.0%), while shared failures on criteria such as Phonotactics, Morphology, and Reflection/Irony expose limitations of the judge itself rather than only weaknesses of the generated corpus.

itations of the judge itself: for several criteria, the model penalizes properties that are also naturally present in therapist-authored materials.

In particular, the judge assigned 0% compliance to both GT and generated texts for Phonotactics and Morphology, despite the fact that such violations are also expected in authentic clinical texts and, in the case of phonotactics, were already shown to be underestimated by the LLM judge relative to manual verification. Likewise, the Reflection/Irony criterion remained near-zero even for GT texts, suggesting that this aspect is not robustly captured by the judge in Russian. Taken together, these results indicate that the LLM-as-a-judge protocol is informative for high-level structural and tonal screening, but should be interpreted cautiously for fine-grained linguistic constraints and abstract semantic properties.

6. Discussion

The results of this study highlight both the transformative potential and the inherent boundaries of using Large Language Models for clinical text generation. A primary finding is the distinct trade-off between therapeutic safety and narrative tension. While the generated texts partially align with the clinical anchor, particularly in readability at the Basic level, they remain structurally simpler than therapist-authored materials in length and syntactic complexity and exhibit a marked “positivity bias” introduced by the safety alignment (RLHF) of the underlying model (Ouyang et al., 2022). Ground-

truth texts, often drawn from classical fables, rely on negative dramatic tension (compound sentiment -0.35) to drive the narrative. In contrast, the generated corpus leans heavily towards positive or neutral resolutions ($+0.22$ to $+0.30$). While this minimizes the risk of triggering emotional distress in patients with post-stroke depression, the lack of conflict in Basic texts (0.74 compliance) suggests that excessive safety constraints may inadvertently sterilize the narrative, potentially reducing patient engagement and motivation.

Furthermore, our analysis reveals a divergence in syntactic density at higher complexity levels. While the generated Basic texts mirrored (and even simplified) the syntactic structure of the clinical anchor, the Intermediate and Advanced levels exhibited subordination rates significantly exceeding the ground truth. This suggests that, without explicit constraints, the model tends to produce a denser narrative syntax than that observed in therapist-authored clinical materials. More broadly, complexity control in LLMs appears to function more as a ceiling than a floor: the model readily simplifies text when strongly constrained, but struggles to preserve the unusually sparse syntax characteristic of therapeutic materials at higher narrative lengths.

A further comparison with the therapist-authored GT texts reinforces this point. When evaluated under the same 18-criterion rubric, GT achieved only a slightly higher aggregate score than the generated corpus (83.3% vs. 80.0%). At the same time, several low-scoring criteria remained problematic for both text types, including Reflection/Irony, Phonotactics, and Morphology. This suggests that part of the observed error signal should be attributed not only to generation failures, but also to limited judge sensitivity for Russian fine-grained linguistic and semantic phenomena.

Finally, the error analysis underscores the “tokenization wall” in prompt-based control. The failure to satisfy phonotactic constraints, combined with the LLM-judge’s inability to accurately assess them, confirms that token-based architectures cannot reliably filter sub-word phonetic features in-context (Sennrich et al., 2016). This necessitates a shift from purely generative pipelines to hybrid neuro-symbolic systems, where the LLM is responsible for semantic coherence and narrative structure, while deterministic post-processors enforce strict phonological and morphological constraints.

7. Conclusions

We present a scalable, controllable pipeline for generating therapeutic reading materials for Russian-speaking patients with aphasia. By formalizing the input space through semantic triplets

and operationalizing clinical progression into discrete complexity regimes, we demonstrated that Large Language Models can produce structurally sound and thematically diverse texts that adhere to strict linguistic constraints. Our evaluation confirms that the generated complexity levels are distinct, with the “Basic” level offering a substantially simpler structural profile, while the “Intermediate” level more closely matches the length and lexical diversity of the therapist-authored clinical anchor.

Future iterations of this work will focus on three key directions. First, we aim to integrate neuro-symbolic post-processing modules to strictly enforce phonotactic rules, addressing the current limitations in consonant cluster control. Second, we will expand the comparative evaluation on therapist-authored Ground Truth texts by increasing the size of the clinical anchor and by comparing judge-based scores with expert human ratings. Finally, and most critically, we will transition from computational evaluation to clinical validation. We plan to deploy the generated corpus in a pilot study with practicing speech therapists (logopedists) and patients to assess the material’s impact on reading comprehension and engagement in a real-world rehabilitation context. This feedback loop will be essential for fine-tuning the generation parameters and moving from a synthetic dataset to a clinically deployable tool.

8. Limitations

Our findings should be interpreted in light of four limitations. (1) Small clinical reference set: the manually curated GT texts serve as a qualitative anchor rather than a representative baseline for Russian aphasia therapy materials; comparisons (e.g., sentiment and syntactic norms) are indicative. Scaling this reference requires digitizing a larger body of clinical manuals. (2) LLM-as-a-Judge bias: despite deterministic decoding and a rubric, judge-based scores may be affected by model bias (e.g., favoring texts aligned with its own distribution) and tokenization limits (as proven by our manual phonotactic verification). (3) Lack of expert validation: while we report a broad set of automated metrics, we have not yet conducted formal blinded expert assessment by practicing speech-language pathologists of the generated texts’ overall therapeutic appropriateness and qualitative utility. (4) No ecological validation: we evaluate computational constraint satisfaction, not therapeutic effectiveness; readability metrics do not guarantee learnability or clinical utility, which also depend on presentation factors (layout, font) and patient-specific deficits.

9. Ethics Statement

This research involves the generation of synthetic texts intended for potential clinical use in aphasia rehabilitation. While these texts are designed to be safe, engaging, and devoid of sensitive or triggering topics, they are experimental and not a replacement for professional clinical judgment. Any deployment of these materials in actual therapy must be done under the strict supervision of qualified speech-language pathologists. Furthermore, the text generation process relied exclusively on generalized archetypes and predefined topics; no real patient data, clinical records, or personally identifiable information was used to train, prompt, or evaluate the models in this study.

10. Acknowledgements

This work was supported by the Basic Research Program at the National Research University Higher School of Economics.

11. Bibliographical References

- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. [Data-driven sentence simplification: Survey and benchmark](#). *Computational Linguistics*, 46(1):135–187.
- Jean-Claude Anscombe and Oswald Ducrot. 1983. *L’argumentation dans la langue*. Pierre Mardaga, Liege.
- Author and Author. 2024. Anonymized. In *Anonymized*, pages 178–180. Anonymized for review.
- Naomi Cocks, Martin Pritchard, Helen Cornish, Nicholas Johnson, and Madeline Cruice. 2013. [A “novel” reading therapy programme for reading difficulties after a subarachnoid haemorrhage](#). *Aphasiology*, 27:509–531.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, et al. 2024. [Gemma: Open models based on gemini research and technology](#).
- Elizabeth Hoover, Ellen Bernstein-Ellis, and Debra Meyerson. 2023. [Using bibliotherapy to rebuild identity for people with aphasia: A book club experience](#). *Journal of Communication Disorders*, 105:106363. Epub 2023 Jul 28. PMID: 37517172.
- William H. Kruskal and W. Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

A. R. Luria. 2008. *Vysshie korkovye funktsii cheloveka [Higher human cortical functions]*. Piter, St. Petersburg. (in Russian).

Philip M. McCarthy. 2005. An assessment of the range and limiting values of measures of lexical diversity. *Department of English, University of Memphis*. Dissertation.

G. Harry McLaughlin. 1969. Smog grading: A new readability formula. *Journal of Reading*, 12(8):639–646.

I. V. Oborneva. 2006. Modified flesch reading ease score for russian texts. *Automatized Analysis of Specialized Texts*, pages 23–28. (in Russian: Avtomatizirovanny analiz spetsializirovannykh tekstov).

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Vladimir Propp. 1968. *Morphology of the Folktale*. University of Texas Press, Austin, TX. Original work published 1928.

Machel Reid, Nikolay Savinov, Denis Teplyashin, et al. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Janet Webster, Julie Morris, and David Howard. 2023. [Reading comprehension in aphasia: the relationship between linguistic performance, personal perspective, and preferences](#). *Aphasiology*, 37(5):785–801.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). arXiv.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. 2024. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A. Representative Generated Examples

To illustrate the controllable generation setup, we provide three representative examples, one for each complexity regime. Each example is annotated with its thematic category, main character, key object, and genre profile. For readability, this appendix includes English translations only; the original Russian versions are available in our repository.

The thematic categories used in the generation framework come from a fixed 12-topic inventory: Vehicles, Wild and Domestic Animals, Home and Garden, Games, Internet and Telecommunications, Arts and Entertainment, Computers and Electronics, AI, Beauty and Fitness, Travel, Sustainability, and Hobbies.

These examples are intended to demonstrate how semantic conditioning and regime-level complexity control interact in practice. They are illustrative and are not used as additional evaluation items.

A.1. Example A: Basic Regime

Metadata. Category: Home and Garden; Main Character: Boy; Key Object: Blue cup; Genre Profile: Real-life incident.

English translation

When he came home, the father asked the children, “What did you do today, children?”

“I washed the dishes,” Olya said proudly.

“And I dried the dishes,” Katya said.

The father turned to his son: “And you, Seryozha?”

“And I picked up the broken pieces,” he replied.

Comment This example illustrates the lowest-complexity regime: a short event structure, an everyday setting, limited narrative development, and high interpretability.

A.2. Example B: Intermediate Regime

Metadata. Category: Home and Garden; Main Character: Girl; Key Object: Yellow flower; Genre Profile: Short narrative.

English translation

A girl named Katya was helping her grandmother in the garden. Her grandmother was teaching her how to care for

plants. Katya especially loved the yellow flower growing by the fence. She watered it every day, carefully loosening the soil around it. One morning, Katya saw that the flower had bloomed. It was so bright and beautiful.

“Grandma, look how beautiful the flower is!” Katya exclaimed.

Her grandmother smiled: “That is the reward for your hard work, dear.”

Comment This example illustrates the intermediate regime through a longer narrative arc, explicit causal progression, and a simple didactic ending while remaining lexically accessible.

A.3. Example C: Advanced Regime

Metadata. Category: Internet and Telecommunications; Main Character: Old Woman; Key Object: Wooden box; Genre Profile: Real-life incident.

English translation

Grandmother Agafya, a woman of about eighty, looked at her granddaughter with distrust.

“This wooden box,” said the granddaughter, “is the Internet. With it, you will be able to talk to your son in America!”

The grandmother carefully took the box and touched its smooth surface. Her granddaughter patiently explained how to use the video call. At first, Agafya was frightened by every click, but when she finally saw her son’s face on the screen, her eyes filled with tears of joy.

“Hello, Mother!” came her son’s voice.

Smiling, the grandmother whispered, “Now this is a miracle!”

Comment This example illustrates the advanced regime through a longer narrative span, more elaborate scene-building, stronger emotional progression, and a clearer multi-step event structure.

An Open-Resource Knowledge Augmentation for Biomedical Lay Summarization

João Pedro Veloso, Evelin Amorim

INESC TEC

{joao.p.veloso, evelin.f.amorim}@inesctec.pt

Abstract

Automatic summarization aims to generate concise versions of texts while retaining relevant information. Summaries can be either extractive, using direct excerpts, or abstractive, rephrasing content to convey the same meaning. Lay summarization applies abstractive techniques to simplify complex texts, such as scientific literature, for broader audiences, thereby promoting public understanding of specialized knowledge. Prior work shows that knowledge augmentation improves lay summarization. Still, biomedical applications often rely on closed resources like the Unified Medical Language System (UMLS), which require expert curation and are costly to scale. We propose a four-step approach that leverages keyword extraction and DBpedia, an open general domain knowledge base, ideal to bridge the gap between expert and lay knowledge. First, we extract keywords from biomedical texts using YAKE!, a well-established unsupervised method. Second, we query DBpedia using these keywords to retrieve relevant concept entries. Third, we construct a graph of concepts for each document based on cosine similarity between DBpedia entries. Finally, we combine each graph with the original abstract to train a summarization model. Our method achieves competitive performance compared to UMLS-based systems in the eLife dataset (ROUGE-1: 58.44 vs. 60.26, ROUGE-L: 43.45 vs. 45.45), demonstrating that open-resource approaches can provide viable alternatives to licensed knowledge bases while maintaining accessibility for resource-constrained organizations.

Keywords: Biomedical Lay Summarization, Knowledge Graphs, DBpedia, Open Resources

1. Introduction

Scientific communication has accelerated rapidly due to the digitization of journals, pre-print servers, and online news (Bornmann and Mutz, 2015; Fraser et al., 2021). This shift has generated an unprecedented volume and velocity of biomedical findings entering the public domain (Landhuis, 2016), necessitating lay summaries to bridge the gap between experts and the general public (Goldstein and Krukowski, 2023). Within Natural Language Processing (NLP), automatic summarization has emerged as a key strategy for condensing complex documents while preserving salient information (Xie et al., 2023; Chaves et al., 2022; Wang et al., 2021). However, lay summarization of biomedical text still faces significant hurdles, including the need to provide sufficient methodological context to prevent model hallucinations and ensure the rigorous preservation of scientific facts (Xiao et al., 2025).

Several approaches have been developed to mitigate hallucinations in biomedical lay summarization. For instance, Goldsack et al. (2023) proposed representing biomedical papers through a combination of text embeddings and concept graphs extracted from the Unified Medical Language System (UMLS)¹. Tang et al. (2023) also used graph-based methods, but of citation papers, to diminish the hallucinations. However, the use of citation papers pro-

vides limited utility for linguistic simplification. Furthermore, reliance on licensed knowledge bases like UMLS restricts the accessibility of these techniques for many organizations. Although UMLS is free for research use, access is not entirely straightforward: it requires registration, periodic license renewal, and typically some form of institutional affiliation. In our experience, this alone can already be a barrier for independent researchers or smaller teams. A more limiting issue is the license itself. UMLS explicitly forbids the redistribution of both the original resource and derived data products. This makes it difficult to share processed datasets or to release systems that depend on it in a fully reproducible way. For example, even when a model can be distributed, anyone attempting to use it must first obtain and configure UMLS independently, which is a non-trivial step. In practice, this makes it difficult to share and reproduce results in line with open research principles. This tension is one of the reasons we explore open-access alternatives, aiming to retain competitive performance while lowering the barrier to entry and enabling broader dissemination. We propose a new technique that employs an open resource for summarization and that presents competitive results with a state-of-the-art licensed technique (Goldsack et al., 2023).

Our proposed technique consists of three stages: (1) automatic keyword extraction from the source text using YAKE! (Campos et al., 2020); (2) retrieval of these keywords from DBpedia and subsequent classification into biomedical and non-biomedical

¹<https://uts.nlm.nih.gov/uts/umls/home>

concepts; and (3) the construction of a concept graph integrated with text embeddings for summary generation. The hypothesis in using keywords is the identification of relevant concepts of a given paper more efficiently, discarding biomedical concepts that are not central to the whole idea of the research. To validate this, we conducted a manual evaluation of 500 extracted keywords (10 per document) to assess their ability to represent the paper's core medical meaning. Also, by using DBPedia, a general domain knowledge base, we aim to bridge the gap between technical and lay vocabulary. To diminish the noise from DBPedia, we apply a biomedical text classification to remove irrelevant concepts. Finally, we connect the concepts by using cosine similarity, and use the built graph to enrich the text representation to generate the lay summary. While UMLS-based approaches have achieved strong performance in biomedical lay summarization, their reliance on licensed resources creates barriers for many organizations, particularly those in low-resource settings or with limited budgets. Our work demonstrates that competitive performance (97% of UMLS-based ROUGE-1 scores) can be achieved using only open resources, thereby democratizing access to biomedical lay summarization technology. To our knowledge, this is the first study demonstrating that open general-domain knowledge bases can substitute licensed biomedical ontologies for lay summarization with minimal performance loss.

Our work aims to answer the following research questions:

RQ1: To what extent can a curated subset of automatically extracted keywords from biomedical papers represent central biomedical concepts?

RQ2: Can a lay summarization technique leveraging open-source knowledge bases compete with approaches using licensed ontologies?

Our contributions are two-fold: (i) we explore the feasibility of using keyword-based representations to preserve medical meaning, validated through a targeted manual assessment; and (ii) we establish that open-source knowledge bases provide a viable and competitive alternative to licensed proprietary databases in the biomedical domain. In the next sections, we describe similar work to ours, emphasizing our contributions, and then describe our method and its results. We release our code to facilitate reproducibility and further research².

²<https://github.com/evelinamorim/openbiolaysumm>

2. Related Work

With the introduction of neural and transformer-based architectures, significant improvements in performance have been made in the summarization of texts. In particular, for biomedical summarization such as (Luo et al., 2023), there has been the exploration of readability-controllable summarization. The authors use transformer models, adapting biomedical content to various levels of readability, which is crucial for bridging expert and lay audiences. In particular, a Longformer-Encoder-Decoder (LED) model is fine-tuned as a baseline model, with additional strategies built as variations upon it. They apply the proposed technique to a new dataset composed of 28,124 articles published in PLOS journals³, where PLOS is a nonprofit open-access publisher covering a wide range of scientific disciplines, including medicine and the life sciences. The authors evaluated their results using ROUGE scores, achieving peaks of 49.87 for ROUGE-1 and 46.02 for ROUGE-L. While our strategy also utilizes the LED architecture, our proposal incorporates an external knowledge base in addition to the source text and summaries used by Luo et al. (2023).

Another approach that incorporates external knowledge is KeBioSum (Xie et al., 2022). The enriched content is based on the PICO framework (Population, Intervention, Comparison, and Outcome) (Huang et al., 2006), which is a structure designed to represent medical knowledge in queries. The method extracts named entities based on the PICO framework using fine-tuned models. Then, the identified elements are combined with the text document to produce the summaries. The authors used three datasets, ORD19 (Wang et al., 2020), PubMed (Cohan et al., 2018), and S2ORC (Lo et al., 2020) to evaluate the effectiveness of their approach. The ROUGE-L values for these datasets were respectively 29.10, 32.28, and 34.08. Despite enriching the content to improve summarization, the exploitation of external concepts is limited only to the identification of relevant terms.

Goldsack et al. (2023), similar to ours, uses the concepts from an external knowledge base, UMLS, and their relations to aid the learning of the biomedical lay summary model. Like our proposal: extraction of relevant concepts, building on the graph of the concepts, and learning the summaries. However, they use a UMLS-based library to identify concepts and their relations. They use two datasets, one composed of papers collected from PLOS journals and another from the eLife journal (Goldsack et al., 2022). The results reported in a combination of the datasets were 48.58 for ROUGE-1 and 45.71 for ROUGE-L. The use of UMLS can limit the usage by organizations, which is not a limitation of

³<https://journals.plos.org/plosone/>

our proposal that uses only open resources (YAKE! and DBpedia) to enrich the input content.

Xie et al. (2024) also employ UMLS to build word and sentence correlation graphs, which are used for topic modeling. The discovery of the topics and summarization are done simultaneously by a neural network, whose inputs are the graphs, the text embeddings, and the gold summary. Like (Xie et al., 2022), in this work, the datasets used are ORD19, PubMed, and S2ORC to evaluate the effectiveness of their approach. The proposed approach achieved ROUGE-1 values between 42.16 and 46.16. Again, the use of a licensed thesaurus limits the use of this approach by organizations.

In the survey by Bhowmik et al. (2025), one challenge highlighted is the need to adopt an open research culture in the biomedical lay summarization task. The reason is that many of the current competitive approaches use the licensed thesaurus UMLS. Our proposal tries to break this barrier with a competitive method based on open-resources.

3. Material and Methods

In this section, we describe the dataset used in our experiments and the main methods we used, including our proposed approach.

3.1. Dataset

Due to limitations of our computational resources, we used a subset of the eLife dataset proposed by Goldsack et al. (2022). The eLife dataset consists of biomedical research articles published in eLife, a peer-reviewed open-access journal in life sciences and biomedicine. Each article includes both the technical scientific abstract and an accompanying “eLife digest”, which is a lay summary written to explain the research to general audiences. Our task is to generate these lay summaries from the technical abstracts, requiring both content selection and linguistic simplification. The selected subset is in our repository, and its statistics are described in Table 1.

Split	# Docs	# Words	Avg. Words
Train	1,582	44.38M	10,212.6 ($\pm 3,479.8$)
Test	241	2.41M	10,009.0 ($\pm 3,324.7$)
Validation	241	2.42M	10,043.5 ($\pm 3,285.9$)

Table 1: Statistics of the eLife dataset subset. (M denotes millions).

3.2. Methods

We evaluated four distinct methods. The first is a BART model (Lewis et al., 2020) fine-tuned on the summaries dataset. The second method appends keywords extracted from YAKE! to the article text and feeds this to a BART fine-tuning process. The third is a reproduction of the approach by Goldsack et al. (2023), hereafter referred to as the UMLS-based method. Finally, we propose OpenBioLaySumm, an approach based on YAKE! (Campos et al., 2020) and DBpedia. In our approach, we make a distinction between keywords and concepts. YAKE! first extracts directly from the abstract keywords— n -gram text spans such as “neuron” or “cell membrane”. We then map these keywords to concepts, i.e., entries in DBpedia obtained via the Lookup API. Each concept is associated with a URI, a textual description, and additional semantic context. For instance, the keyword “neuron” is mapped to the DBpedia resource <http://dbpedia.org/resource/Neuron>, along with its natural language description.

We implement the doc-enhance architecture from Goldsack et al. (2023), which combines textual and graph-based representations through a multi-component encoder-decoder framework. The pipeline consists of four main steps (Figure 1):

1. **Concept Extraction:** In the UMLS-based variant of our approach, biomedical concepts are identified through mappings to the UMLS Metathesaurus. Goldsack et al. (2023) originally relied on MetaMap for this purpose. However, following recent changes in the UMLS distribution (notably the discontinuation of the MetaMap NLP Content View in the 2025AB release⁴), we instead implement concept recognition using QuickUMLS (Soldaini and Goharian, 2016), which enables approximate string matching over the Metathesaurus and can be readily integrated with custom subsets. By contrast, OpenBioLaySumm does not directly rely on UMLS mappings for initial concept detection. Instead, it applies YAKE! (Campos et al., 2020) to extract the top-ranked keywords from each biomedical abstract. These keywords are subsequently used to query the DBpedia Lookup API⁵, which returns a ranked list of candidate entities for each query term, yielding up to 100 entity candidates per document. Because DBpedia includes entities from multiple domains, an additional filtering step is required to retain only biomedical concepts. To this end, candidate entities are automatically classified as biomedical or

⁴https://www.nlm.nih.gov/pubs/techbull/nd25/nd25_umls_release_ab.html

⁵<https://lookup.dbpedia.org/>

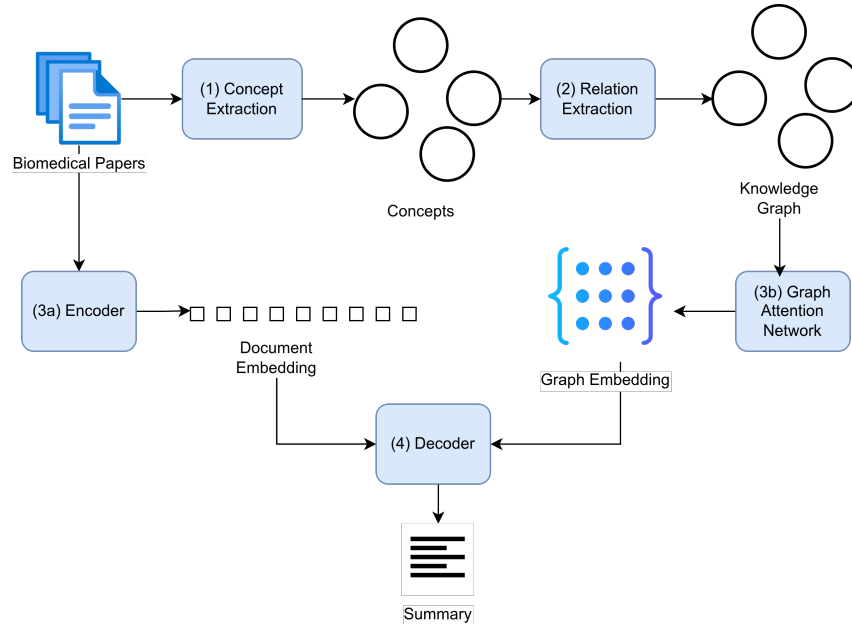


Figure 1: Pipeline for both summarization approaches (UMLS-based and OpenBioLaySumm), following the doc-enhance architecture from [Goldsack et al. \(2023\)](#). Both approaches: (1) extract concepts from biomedical abstracts, (2) construct concept graphs: the UMLS-based approach uses relations provided by the UMLS Metathesaurus, while OpenBioLaySumm computes edge weights via SciBERT cosine similarity > 0.7 , (3a) encode the document with LED, (3b) process the graph with a 3-layer GAT, then concatenate both representations through an additional encoder layer, and (4) generate the lay summary with the LED decoder.

non-biomedical using a zero-shot large language model based on their textual descriptions. The results were classified using the description of the entry and DeepSeek-R1:7b as biomedical or non-biomedical concepts with the prompt: “Classify if the following term is related to the biomedical/biological domain: concept description. Answer yes if it’s related to biomedical/biological, no if not.”. Concepts not assigned to the biomedical domain are discarded prior to graph construction. A manual validation of these LLM classifications was performed by one of the authors to ensure quality (described in Section 4.1).

2. **Relation Extraction:** The two approaches differ mainly in how the concept graphs are constructed. In the UMLS-based setting, we rely on the semantic relations already defined in the UMLS Metathesaurus (e.g., *is_a*, *part_of*, *treats*), which are extracted together with the concepts using QuickUMLS. This results in graphs that reflect curated, domain-specific knowledge. By contrast, OpenBioLaySumm does not assume a fixed set of relations. Instead, relationships are inferred dynamically from semantic similarity. To do this, we encode the DBpedia description of each concept using SciBERT ([Beltagy et al., 2019](#)), truncating inputs to 250 tokens. We then apply mean

pooling over the token embeddings to obtain a 768-dimensional vector for each concept. To make this step more efficient, we reduce the representations to 50 dimensions using PCA fitted on the training data. Once the representations are obtained, we compute cosine similarity between all pairs of concepts within a document. Edges are added between concepts whose similarity exceeds 0.7, yielding a relatively sparse graph in which each concept is linked to a set of semantically related neighbours rather than to a predefined relation type.

3. **Graph and Document Encoding:** Following the doc-enhance architecture, we encode information from both textual and structured modalities. The document text is processed using the Longformer-Encoder-Decoder (LED) model ([Beltagy et al., 2020](#)), which produces document-level representations suitable for long biomedical inputs (approximately 10,000 words on average in our dataset) (Figure 1, 3a). In parallel, the corresponding concept graph is encoded with a 3-layer Graph Attention Network (GAT) (hidden size 768)(Figure 1, 3b). ELU activations are applied in the first two layers, followed by a linear projection in the final layer. To combine both modalities, we concatenate the document embedding obtained from

LED with the graph-level representation produced by the GAT. The resulting representation is then passed through an additional encoder layer, allowing interactions between the textual and graph-based features prior to decoding.

4. **Summary Generation:** We use the Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) as our base model, where the encoder processes the concatenated document and graph representations, and the decoder generates the lay summary. In the training, the summary is fed to the decoder to learn such an output.

Our implementation builds on the publicly available code provided by ~ Goldsack et al. (2023), while preserving the original doc-enhance architecture. We made the following changes to support our experimental setup: (1) we replaced MetaMap with QuickUMLS for UMLS concept extraction following recent changes in UMLS content views, (2) we incorporated a concept extraction step based on YAKE! and DBpedia for OpenBioLaySumm, and (3) we included a SciBERT-based similarity module to construct the concept graph during training.

4. Experiments and Results

We did three experiments regarding our method: (1) the extraction of the concepts, (2) the classification of the concepts in biomedical entities, and (3) the lay summarization process. Next, we detail each one of them and its results.

4.1. Extraction of the Concepts

This stage was performed by the UMLS-based method and OpenBioLaySumm. For the UMLS-based method, we adapted the code made available for the paper (Goldsack et al., 2023) in a GitHub repository⁶. The adaptation is because the tool used by the authors, MetaMap, was not maintained anymore⁷. Thus, we use an alternative, the QuickUMLS library which presents the same functionality of extracting UMLS concepts (Soldaini and Goharian, 2016). In addition to the extraction, this library also identifies the class of the concept and the relations between the concepts. Therefore, the next stage, classification of concepts, is not necessary for the UMLS-based method.

For OpenBioLaySumm, we exploit the keyword extraction called Yake! proposed by Campos et al.

(2020), which uses a statistical technique to identify the main keywords in a given text. We apply Yake! for each document with the number of terms (top) set to 10, n-grams of up to 2 words ($n = 2$), which means only unigrams and bigrams, and a variable deduplication limit (dedupLim) set to 0.5. Deduplication merges similar keywords to avoid redundancy, and a value of 0.5 merges similar keywords with 50% similarity according to the Levenshtein distance. Examples of biomedical keywords extracted from the test set are: “neck region”, “cell membrane”, “cell growth”, and “endocytic events”. Some non-biomedical keywords are also extracted, but this is not relevant to the generation of the lay summary; that is why we discard these concepts using an automatic classification step, which is described in Section 4.2. Examples of non-biomedical words in the test set are: “figure supplement”, “Supplementary file”, “Time constant”, and “label set”.

To assess whether Yake! is able to extract terms that express biomedical concepts, one of the authors of this paper evaluated the extracted keywords. For this evaluation, we randomly selected 50 papers from the test set, and the 10 keywords extracted for each paper were evaluated as biomedical and non-biomedical. The total of evaluated keywords was 500 terms from the test set. In this set, 151 keywords were not considered biomedical, while 349 keywords were considered biomedical. Thus, 70% of the keywords extracted by Yake! were biomedical.

4.2. Classification of the Concepts

As described in Section 3.2, UMLS-based methods do not require domain classification since the UMLS Metathesaurus is curated by specialists. However, OpenBioLaySumm requires filtering since YAKE! extracts keywords from all domains without biomedical discrimination. This section presents the experimental validation of our zero-shot LLM classification approach.

To assess the capabilities of the model to classify a given concept in a zero-shot manner into biomedical or non-biomedical, one of the authors evaluated the output of the model. Similarly to the keywords evaluation, 10 keywords and their respective DBpedia entries of a sample of 50 documents were analyzed, for a total of 500 DBpedia entries. Cohen’s Kappa agreement between the LLM classifier and a human evaluator was computed per document, and the average of the values was 0.3324. Half of the agreements were poor ($22, \kappa < 0.20$), a few presented a slight ($4, 0.20 \leq \kappa < 0.40$), and the rest were moderate to almost perfect ($24, \kappa \geq 0.40$). In a more detailed analysis, it was possible to observe that the LLM classified a lot of the concepts as biomedical, probably because it is a general domain LLM, and it needed examples to understand

⁶https://github.com/TGoldsack1/Enhancing_Biomedical_Lay_Summarisation_with_External_Knowledge_Graphs

⁷<https://wayback.archive-it.org/7867/20241213191952/https://lhncbc.nlm.nih.gov/ii/tools/MetaMap.html>

what a biomedical concept is. However, since the majority of the concepts extracted are biomedical (70%), the influence of the mistakes of the classification is probably small in the summarization result. We leave as future work the application of a few-shot approach at this stage of the proposal.

4.3. Lay Summarization

For the OpenBioLaySumm, it is necessary to relate each concept to the others. In this stage, we use SciBERT and compute the cosine similarity between the concepts. Concept pairs with cosine similarity above 0.7 are connected by edges in the graph. The representation of the graph is learned by the Graph Attention Network (GAT) and concatenated with the LED text embedding of the document, and fed to a Decoder, which is then responsible for producing the summary. We trained all four models (BART baseline, BART+YAKE!, UMLS reproduction, and OpenBioLaySumm) using a learning rate $2e-6$, and a batch size 2-4. All models used identical decoding parameters: beam search with 4 beams, minimum length 100 tokens, length penalty 2.0, and no-repeat-ngram size 3. However, due to limitations in our computational resources, the UMLS (Goldsack) strategy was trained only with 5 epochs, while BART, BART+YAKE!, and OpenBioLaySumm (YAKE!+DBPedia) were run with 20 epochs since they all required less GPU memory than the UMLS approach. All models were trained under an identical four-hour GPU budget on an NVIDIA A100 (40GB). Due to the higher memory footprint and preprocessing overhead of the UMLS-based pipeline, it completed five training epochs within this time constraint, while the other models completed approximately twenty epochs. Under the four-hour fixed GPU budget, the UMLS-based model validation performance peaked at the first epoch and did not show consistent improvement in subsequent epochs, suggesting early convergence under the given setup. Early stopping was applied uniformly across models.

Table 2 depicts the results of the four approaches. Observe that among the open-based approaches (BART, BART+YAKE!, and OpenBioLaySumm), the best results in three metrics are from OpenBioLaySumm. Another interesting observation is that BART+YAKE! improved over the baseline, demonstrating that keyword augmentation alone provides value even without graph construction. The ROUGE-Lsum for the BART+YAKE! approach is also the second-best result, which shows a better summary-level sequential coherence. This is expected: prepending keywords maintains a simple linear structure, while our GAT-encoded graph introduces richer but more complex concept relationships that may occasionally reorder information. However, the gain in the three other metrics

achieved by OpenBioLaySumm is more significant compared to the gain of ROUGE-Lsum.

The lay summarization task requires simplification of the text. Hence, readability measurements are a usual way to inspect the level of simplification of the results. Figure 2 shows a comparison between the gold summaries (reference) and three of the approaches tested. The Flesch score is a readability metric based on which the higher the value, the simpler the text. All the methods were able to simplify the text beyond the gold summaries. The UMLS (Goldsack) approach presented the best simplification results. Regarding the Coleman-Liau and Dale-Chall scores, the smaller the value, the simpler the text. These metrics indicate the same; the UMLS (Goldsack) presented the lowest values, while OpenBioLaySumm (YAKE!+DBPedia) presented the second lowest ones. The remaining measures, average sentence length, average word length, and type-token ratio, shows similar results. One hypothesis for OpenBioLaySumm (YAKE!+DBPedia) presents this type of result, which is the type of text that DBPedia entries have compared to the text of the UMLS definitions. As a general domain knowledge base, DBPedia probably adds a more diversified vocabulary to the final summary compared to a specialist-curated knowledge base, such as UMLS.

In this section, we presented quantitative measures of the quality of the summaries. However, we acknowledge ROUGE's limitations: it measures lexical overlap rather than semantic fidelity, factual correctness, or accessibility for lay readers. Qualitative analysis is essential in those cases, although it would require expensive expert work. To mitigate the quantitative metrics limitations, and give a more qualitative view of the results we provide in the section, an analysis through two illustrative examples: a success case where OpenBioLaySumm (YAKE!+DBpedia) outperforms the UMLS (Goldsack) approach, and a failure case where it achieves significantly lower BLEU scores.

5. Discussion

We will divide our discussion into two parts. The first will discuss the quantitative results we presented in the last section, and the second part will present two examples and qualitatively discuss the UMLS-based approach (Goldsack) and the OpenBioLaySumm (YAKE! + DBPedia). The goal is to deepen the understanding of the enrichment role in the lay summarization task.

Quantitative results. We evaluated the three main stages of our pipeline: the keyword extraction, the classification of concepts in the biomedical domain, and the final lay summarization. The first evaluation helps to answer our first research question:

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
BART Baseline	45.84	11.85	20.92	42.55
BART+YAKE!	47.53	12.97	21.63	<u>44.04</u>
OpenBioLaySumm (YAKE!+DBpedia)	<u>58.44</u>	<u>20.65</u>	<u>43.45</u>	43.47
UMLS (Goldsack et al.) [†]	60.26	22.55	45.45	45.44

Table 2: ROUGE scores on the test split. [†]Our reproduction using QuickUMLS and 5 training epochs. [Goldsack et al. \(2023\)](#) reported ROUGE-1: 48.58 and ROUGE-L: 45.71 on PLOS+eLife combined dataset using MetaMap and full training. Direct comparison is limited due to dataset and implementation differences.

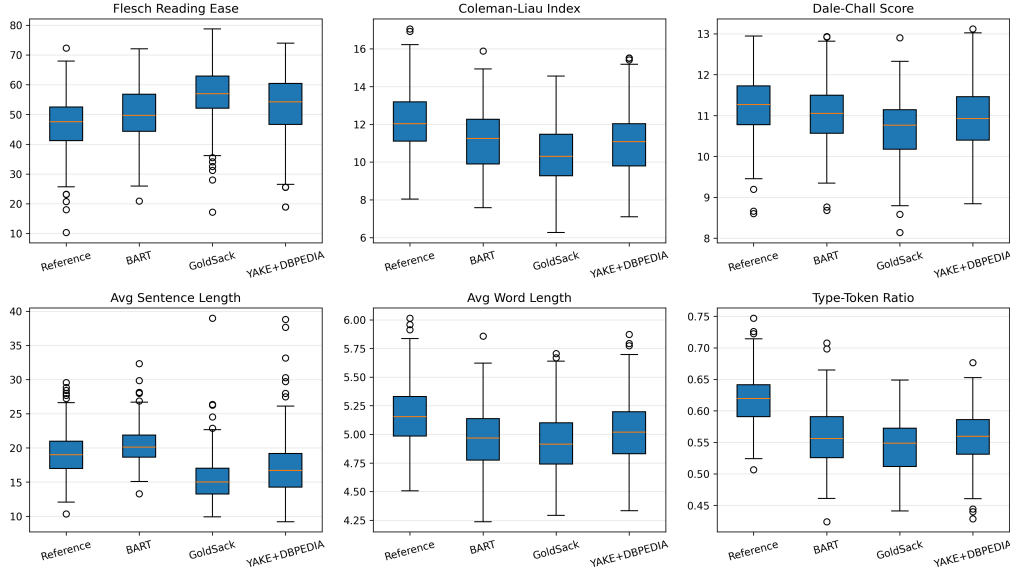


Figure 2: Readability measurements for Gold Summary, BART (baseline), OpenBioLaySumm(YAKE!+DBPedia), and UMLS (Goldsack)

RQ1) To what extent can a curated subset of automatically extracted keywords from biomedical articles represent central biomedical concepts? The results of the assessment indicate that Yake! effectively retrieves biomedical meaning from biomedical papers. In an evaluation of 50 random documents from our test set, we assessed that approximately 70% of the keywords extracted from the biomedical papers are domain-related. The minimum of biomedical terms in a document was 2, and the maximum was 10. Despite the minimum of 2 keywords, the average was 6.96 ± 1.79 and the mode was 8. These numbers suggest that Yake! can extract medical terms from the papers. This is an intuitive result since scientific papers usually follow a writing template, and the authors usually convey only the essential information to the readers, avoiding excess that can confuse.

The LLM classification results presented, most of the time, low agreement scores with the human evaluation. In this case, we also evaluated 50 documents, and 26 presented Cohen’s kappa lower than 0.40. An inspection of the LLM results revealed that the LLM alone is not capable of being

aware of what a biomedical concept is. We use a zero-shot approach, which can be the reason for such a low performance. The LLM classifier identified 444 of the retrieved concepts as biomedical, while the human evaluator identified only 349 of the 500 concepts. One possible solution for improvement is to develop a domain classifier, which could take advantage of the biomedical words to sieve the relevant entries. An only-encoder model could be fine-tuned to distinguish biomedical and non-biomedical terms based on their collected DBPedia entries. A domain-specific classifier could sieve the concepts more effectively, reducing the noise of out-of-domain concepts in the GAT. This could improve the results for an approach based on the YAKE! and DBPedia.

Finally, the results of the summarization show that the OpenBioLaySumm achieved 96.97%, 91.57%, 95.59%, and 95.96% of the performance in the respective ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum scores of the UMLS-based approach (Goldsack). In a close look at the readability metrics, we can observe that the average sentence length of the OpenBioLaySumm is greater

than that of the UMLS one (Goldsack). DBPedia uses an encyclopedic style in the writing, and compared to a more specialized thesaurus like UMLS, it probably contains more lengthy sentences and less objective explanations. The results, therefore, are a reflection of the DBPedia writing style. The same influence of the DBPedia style is possible to observe in the Type-Token Ratio (TTR) boxplot. This metric measures the number of unique tokens divided by the total number of tokens. The boxplot shows that the average of TTR is slightly higher than the BART approach, indicating that DBPedia introduced a more diverse vocabulary. The noise from the non-biomedical entries of the DBPedia would probably even worsen this effect. We also computed the BLEU scores for the summaries. Figure 3 presents the overall BLEU scores, in which the UMLS (Goldsack) presented a slightly superior advantage. The analysis of these results answers our second research question, **RQ2**) Can a lay summarization technique leveraging open-source knowledge bases compete with approaches using licensed ontologies? The answer to this question is positive since our preliminary results with an open-resource approach reached between 91.57% and 96.97% in the ROUGE metrics.

Qualitative results. Regarding the extracted keywords, it is possible to observe that some of the identified terms that are not biomedical are very repetitive. For instance, the word “figure” appears in 72 of the extracted keywords, and “supplement” appears in 43. Differently, the most frequent biomedical keywords present much lower frequency values. For instance, “cells”, “dna” and “cell” appear in 18, 16, and 15 keywords. This is probably because the articles in this particular dataset cover a broad number of topics within the biomedical domain, and the discovered keywords are meaningful for summarization.

The LLM classification, despite its poor results, was able to filter a very common non-biomedical keyword, “figure supplement”. However, the LLM made simple mistakes like classifying the concept *Politician*⁸ as biomedical. This type of error caused a lot of noise in the final result.

For the summarization results, we will look at two examples. The first example is depicted in Table 3, the OpenBioLaySumm shows better performance than the UMLS (Goldsack) approach. Observe that in this example, both have low BLEU values, but OpenBioLaySumm presented a better phrasing, and the concepts in this abstract (“electrical impulse”, “vesicles”, “membrane”, etc) are very different, which favors the DBPedia approach.

We examined failure cases to understand when graph-based augmentation degrades performance.

⁸<http://dbpedia.org/resource/Politician>

Success Case: Neuroscience Abstract (Synaptic Transmission)

BLEU_UMLS = 8.804; BLEU_Ours = 15.906 (+81%)

UMLS:	Neurons	communi-
cate	with	each another ...
a electrical signal is between a nerve ...		vesi-
cles	are travel to each on	the membrane...
OpenBioLaySumm:	Neurons	
communicate	with	one another ...
a electrical impulse reaches from a neuron ...		
vesicles	are leave to receptors on	the surface...
Green = Better phrasing; Red = Errors in UMLS		

Table 3: OpenBioLaySumm achieves 81% higher BLEU by better capturing key biological relationships (receptors, synapses, vesicles) through DBpedia’s general-domain knowledge graph.

Table 4 shows an example where OpenBioLaySumm achieves only 34.34% of UMLS’s BLEU score (7.016 vs. 20.432) on a text about Notch signaling. The failure exhibits two issues. First, both methods struggle to correctly tokenize and preserve protein family names (Delta, Serrate, Jagged), producing fragments like “Seragged”, “Lagagged”, and repeated single-letter tokens “D D D”. Second, OpenBioLaySumm’s graph structure amplifies these errors: the high semantic similarity between protein names (Notch, DLL1, DLL4, Jagged) creates densely connected subgraphs. This causes the GAT attention mechanism to cycle through similar fragmented tokens, leading to severe repetition patterns like “of of both of some of” that further degrade BLEU scores. Despite this type of case, our method’s overall strong performance (ROUGE-1: 58.44, 96.97% of UMLS) indicates this is a bounded problem affecting a small subset of highly technical abstracts.

6. Conclusion

In this work, we propose an open-resource approach to lay summarization of biomedical papers. We showed that the performance of the proposed approach is competitive, and information extraction can be achieved in different ways other than exclusively named entities, like YAKE!. Finally, we showed examples of the test set that demonstrate the strengths and weaknesses of our method.

However, this work has several limitations. One of them is that we only tested one LLM for the classification, and nowadays there is a diverse set of models that could be tested. Another one is that, due to constraints in computational resources, we did not

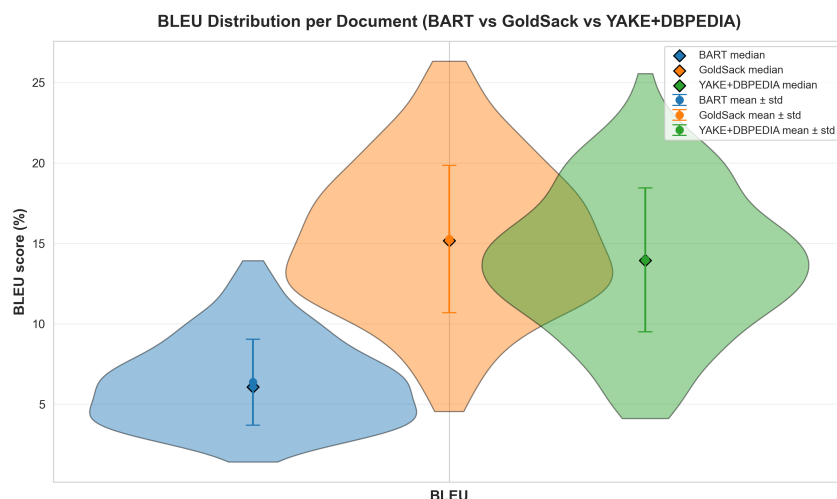


Figure 3: The BLEU scores for the UMLS (Goldsack) and the OpenBioLaySumm (YAKE!+DBPedia)

UMLS (Goldsack) BLEU = 20.432	OpenBioLaySumm BLEU = 7.016 (-65.66%)
Protein name fragmentation:	
receptors lig Seragged ligands bind...	to a on lig- Lagagged ligands bind...
Mammals have Not of Not ligands, D of of Jagged...	Theals have Not of lig ligands, D that of Jagged...
DLL1 and DLL4 are activate in in cells cells , of, D D	DLL1 and DLL4 are bind on on the cells ,
some tissues D D D ...	of of both of some of have D type...
Structural errors:	
Onealing are between and forth	cells Oneals from from and forth
on on and off ...	as and on on and off ...

Table 4: Failure case showing severe degradation with protein-heavy terminology. Both methods fragment protein names (Serrate/Jagged → "Seragged"/"Lagagged", Delta → "D"), but OpenBioLaySumm shows more repetition ("of of both of some of" , 66% BLEU drop). The dense graph of similar protein concepts amplifies tokenization errors through attention cycles.

test several different parameters for the model, including YAKE! parameters. The prompt is also very simple, and a few-shot approach could improve the results. Also, the evaluation was conducted by only one person, who is also one of the authors of the paper. An additional human evaluation would strengthen the conclusions of the proposed approach. Our approach struggles with abstracts containing many similar, interconnected technical entities (e.g., protein families with multiple related members). The graph structure, while beneficial for most documents, can amplify confusion in such cases by creating densely connected concept clusters that lead to repetitive generation. Also, due to computational constraints, we reduced SciBERT embeddings from 768 to 50 dimensions via PCA. While this enabled processing within our GPU budget, we did not systematically evaluate the impact of this reduction on final performance. Future work should explore the trade-off between dimensional-

ity, computational cost, and summarization quality. Finally, the classification could benefit from a standard approach with a BERT-based classifier trained on general data combined with biomedical data. This could be an efficient and effective approach compared to an in-context learning approach based on LLMs.

Future work should explore adaptive graph pruning that reduces edge density when concepts exceed a similarity threshold in localized graph regions. Also, we plan an ablation study with the used parameters and do a comparison between our GAT-based approach against simpler aggregation methods (e.g., mean-pooled concept embeddings) to isolate the contribution of graph-based reasoning. This would clarify whether performance gains stem from DBpedia's knowledge content or from the graph structure itself. Finally, we plan to conduct a more extensive evaluation of the results with biomedical experts.

7. Acknowledgements

We thank Tomas Goldsack for making the code from [Goldsack et al. \(2023\)](#) publicly available. Our implementation builds upon their doc-enhance architecture, with modifications for QuickUMLS integration and the YAKE!+DBpedia pipeline. The authors acknowledge the computational resources provided by the Deucalion supercomputer at the Minho Advanced Computing Centre (MACC), University of Minho, Guimarães, Portugal, through projects 2025.00013.AlvLAB.DEUCALION and 2025.00017.AlvLAB.DEUCALION. Deucalion is funded by the EuroHPC Joint Undertaking and the Portuguese Foundation for Science and Technology (FCT). Also, this publication is based upon work from COST Action CA23147 GOBLIN – Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, <https://www.cost.eu>).

8. Bibliographical References

- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Souramita Bhowmik, Anupam Jamatia, Dwijen Rudrapal, and Kunal Chakma. 2025. Biomedical lay summarization: Research progress and challenges. In *2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*, pages 464–471. IEEE.
- Lutz Bornmann and Rüdiger Mutz. 2015. [Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references](#). *J. Assoc. Inf. Sci. Technol.*, 66(11):2215–2222.
- Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.
- Andrea Chaves, Cyrille Kesiku, and Begonya Garcia-Zapirain. 2022. [Automatic Text Summarization of Biomedical Text Data: A Systematic Review](#). *Information*, 13(8):393. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.
- 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. *arXiv preprint arXiv:1804.05685*.
- Nicholas Fraser, Liam Brierley, Gautam Dey, Jessica K. Polka, Máté Pálffy, Federico Nanni, and Jonathon Alexis Coates. 2021. [The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape](#). *PLOS Biology*, 19(4):1–28. Publisher: Public Library of Science.
- Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. [Making science simple: Corpora for the lay summarisation of scientific literature](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tomas Goldsack, Zhihao Zhang, Chen Tang, Carolina Scarton, and Chenghua Lin. 2023. Enhancing biomedical lay summarisation with external knowledge graphs. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Carly M Goldstein and Rebecca A Krukowski. 2023. The importance of lay summaries for improving science communication. *Annals of Behavioral Medicine*, 57(7):509–510.
- Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. 2006. Evaluation of pico as a knowledge representation for clinical questions. In *AMIA annual symposium proceedings*, volume 2006, page 359.
- Esther Landhuis. 2016. [Scientific literature: Information overload](#). *Nature*, 535(7612):457–458.
- 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.
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S 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.
- Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. [Readability Controllable Biomedical](#)

[Document Summarization](#). ArXiv:2210.04705 [cs].

Luca Soldaini and Nazli Goharian. 2016. Quick-umls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Chen Tang, Shun Wang, Tomas Goldsack, and Chenghua Lin. 2023. Improving biomedical abstractive summarisation with knowledge aggregation from citation papers. *arXiv preprint arXiv:2310.15684*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, et al. 2020. Cord-19: The covid-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Mengqian Wang, Manhua Wang, Fei Yu, Yue Yang, Jennifer Walker, and Javed Mostafa. 2021. [A systematic review of automatic text summarization for biomedical literature and EHRs](#). *Journal of the American Medical Informatics Association : JAMIA*, 28(10):2287–2297.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William K Cheung, et al. 2025. Overview of the biolay-summ 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.

Qianqian Xie, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. 2022. [Pre-trained language models with domain knowledge for biomedical extractive summarization](#). *Knowledge-Based Systems*, 252:109460.

Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. 2023. [A Survey for Biomedical Text Summarization: From Pre-trained to Large Language Models](#). ArXiv:2304.08763 [cs].

Qianqian Xie, Prayag Tiwari, and Sophia Ananiadou. 2024. [Knowledge-Enhanced Graph Topic Transformer for Explainable Biomedical Text Summarization](#). *IEEE Journal of Biomedical and Health Informatics*, 28(4):1836–1847.

TabMedQA: From Structured Data to Question-Answer Datasets in Early Clinical Decision-Making

Gabriel Iturra-Bocaz, Petra Galuščáková, Sol Vedde, Alvaro Fernandez-Quilez

University of Stavanger
{gabriel.e.iturrabocaz, petra.galuscakova, alvaro.f.quilez}@uis.no
Stavanger University Hospital
s.gedde@sus.no

Abstract

The rising adoption of Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) in clinical general practice demands datasets that capture realistic early-stage clinical decision-making, where experts must decide on follow-up actions based on sparse, structured patient data. Existing medical Question-Answering (QA) resources primarily address post-diagnostic or specialist settings and rarely reflect how General Practitioners (GPs) document and justify early decisions based on clinical observations from Electronic Health Records (EHRs) and grounded on clinical guidelines. We present TabMedQA, a framework for synthesizing QA collections that emulate how GPs formulate and document decisions in encounter notes during early patient assessments. TabMedQA leverages instruction-tuned LLMs, guided by disease-specific clinical guidelines, to generate full encounter notes composed of a guideline-grounded justification and a corresponding follow-up recommendation directly from structured EHR inputs. The framework further supports RAG-based evaluation, simulating how GPs might consult previous patient encounters to inform new consultations. We demonstrate the application and resulting resource use of TabMedQA on prostate cancer using the publicly available PI-CAI collection and release the resulting PI-CAI QA collection, resource generation templates, and TabMedQA code. To the best of our knowledge, TabMedQA provides the first open framework for creating guideline-grounded, EHR-based QA collections that enable the generation and holistic evaluation of LLM-produced clinical encounter notes, bridging decision-making accuracy with clinical encounter quality in general practice.

Keywords: Medical QA Datasets, Tabular Data, EHR, Retrieval Augmented Generation, decision-making

1. Introduction

General Practitioners (GPs) operate at the earliest stage of clinical decision-making, where they have to interpret sparse and highly heterogeneous patient information to propose a clinically grounded and justified course of action (Starfield et al., 2005). For this purpose, they typically rely on limited structured data from Electronic Health Records (EHRs) such as demographics, risk factors and medical test screening results (Crosson et al., 2005). By the end of each patient consultation, the GP must document a medical encounter note that includes (i) a recommendation regarding patient follow-up and (ii) a justification of their decision grounded in clinical observations and established medical guidelines (Europe, 2002). These encounter notes serve not only as formal documentation of the consultation, but also as a legal and ethical safeguard ensuring traceability of decisions and transparency towards patients (Rose and Joshi, 2018; Sittig and Singh, 2011).

Large Language Models (LLMs) equipped with Retrieval Augmented Generation (RAG) are increasingly integrated into clinical workflows of GPs to assist with documentation, information retrieval and decision support (Gaber et al., 2025; Lu et al., 2024). Such systems can access collections of previous patient encounters and their medical notes,

allowing them to ground their responses in comparable cases. Within this context, the process of generating a medical encounter note from EHR data can be naturally viewed as a question-answering (QA) task (Figure 1). Here, the GP acts as the user who formulates a question to an LLM system that has access to the patient’s structured information, and, optionally, similar historical cases, requesting an answer in the form of an encounter note that outlines an appropriate follow-up recommendation together with a justification consistent with clinical guidelines. However, in spite of the relevance of the scenario in real-world clinical practice, collections supporting the systematic development and evaluation of LLM with RAG emulating GP-LLM interaction to produce clinical encounter notes remain scarce (Shool et al., 2025; Elgedawy et al., 2024; Gomez-Cabello et al., 2024).

Existing clinical QA datasets primarily address specialist or post-diagnostic settings, relying on the availability of medical notes written by experts or extracting answers from complete clinical narratives (Jin et al., 2021; Singhal et al., 2025; Kim et al., 2024). In both cases, these resources are difficult to obtain at scale because of privacy restrictions, the high cost of expert annotation and the variability in clinical documentation practices (Williamson and Prybutok, 2024; Kitamura et al.,

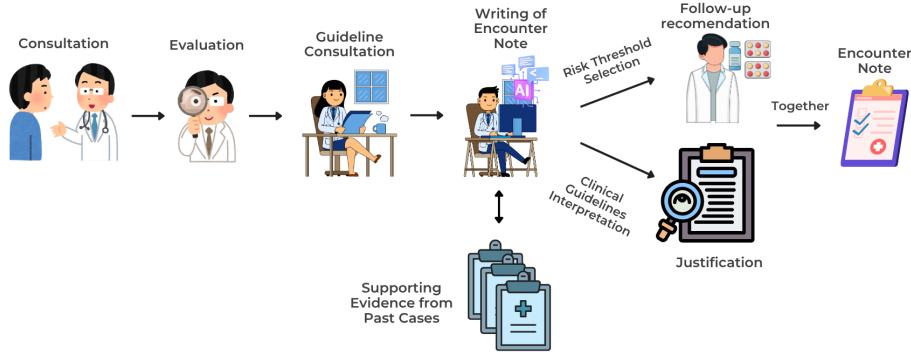


Figure 1: Overview of a GP’s consultation workflow, from patient evaluation to the creation of an encounter note containing a follow-up recommendation and its clinical justification.

2024; Cohen et al., 2019). Furthermore, existing clinical QA collections provide brief factual answers or diagnostic labels rather than guideline-grounded encounter notes that explicitly articulate both the justification behind a clinical decision and the corresponding decision (Kim et al., 2024). As a result, current datasets and benchmarks provide limited support for studying how can LLMs generate guideline-consistent encounter notes from EHR structured data, or how to perform realistic and holistic evaluations of encounter notes including decision-making and overall clinical encounter note quality (Bardhan et al., 2024).

To address these gaps, we introduce **TabMedQA**, a framework to construct medical QA collections that simulate how GPs could interact with RAG-equipped LLMs during initial patient evaluations to produce medical encounter notes. The framework leverages instruction-tuned LLMs guided by disease-specific clinical guidelines to generate question-answer pairs directly from structured EHRs, where each answer corresponds to an encounter note combining a clinical-guideline grounded justification and a follow-up recommendation (Figure 2).

At inference time, resources generated with TabMedQA support RAG-based evaluation, simulating how a GP might consult an automated system that retrieves previous patient encounter notes to inform new follow-up recommendations and justifications (Figure 1). We demonstrate its use through one study case to depict the application of TabMedQA framework and resulting QA collection in a clinically realistic early decision-making scenario. Our main contributions can be summarized as follows:

- We introduce **TabMedQA**, a framework for constructing medical QA collections that simulate early general-practice consultations (Figure 1). TabMedQA synthesizes encounter notes with both clinical justifications and follow-up

recommendations from structured EHR data, guided by disease-specific clinical guidelines and instructed LLMs.

- We show how TabMedQA-generated QA collections can be used in clinical real-world scenarios in which an LLM accesses **a bank of previous patient cases** to inform recommendations and justifications for a given new patient EHR and visit. This mirrors how GPs could interact with LLM-based systems in clinical practice during new patient visits.
- We release a QA collection focuses on prostate cancer (PI-CAI QA), together with the corresponding generation templates and TabMedQA code to support future research on early-stage justified clinical-decision making¹.

The remainder of the paper is organized as follows: Section 2 reviews relevant literature, Section 3 presents the methodological approach to construct the **TabMedQA** framework, Section 4 describes the resulting QA collection obtained by applying TabMedQA to a clinical use case and provides guidance on how the resource can be leveraged at inference time to simulate GP–LLM interactions, Following, Section 5 details experiments evaluating these interactions in RAG-based settings where prior encounter notes are used as a support to synthesize new encounter notes. Section 6 discusses results and findings, whilst Section 7 concludes with perspectives for future work. Finally, Section 8 we discuss the limitation our proposed TabMedQA framework.

2. Related Work

The rapid adoption and integration of LLMs equipped with RAG in clinical practice has increased the research community interest in under-

¹<https://github.com/iai-group/cl4health-tabmedqa/>

standing their impact in decision support and automatic documentation (Eguia et al., 2024; Gaber et al., 2025). However, most existing work in this area has focused on specialist or post-diagnostic clinical scenarios (Ke et al., 2025), with limited attention to how they could support early, GP settings where data are primarily structured and decisions must be justified according to clinical guidelines (Gomez-Cabello et al., 2024; Korom et al., 2025). In this section, we review related work along two complementary directions: (i) the use of RAG systems in healthcare applications and (ii) the development of QA datasets for evaluating clinical LLMs. Together, they highlight the absence of resources that combine structured EHR data, guideline-grounded justifications and retrieval-based access to prior patient cases emulating realistic GP workflows (Figure 1).

2.1. RAG in Healthcare

In the healthcare domain, RAG systems have been explored for a variety of applications, including the interpretation of clinical guidelines, clinical decision-support, and diagnostic assistance (Amugongo et al., 2025; Ge et al., 2024; Rau et al., 2024). For example, Ge et al. (2024) introduced LiVersa, a liver disease-specific chat bot that leverages RAG systems to perform clinical answering for hepatology questions. Similarly, Rau et al. (2024) developed a RAG-based chat bot designed for gastrointestinal imaging, which achieved significantly higher accuracy than generic GPT-4 models. Recent studies have also applied RAG to medical QA and literature-grounded reasoning tasks, retrieving evidence from textual biomedical sources rather than patient data (Singhal et al., 2025; Zhang et al., 2025). While these systems demonstrate the potential of RAG in specialized clinical contexts, they primarily rely on retrieving textual medical knowledge to support post-diagnostic decision-making. In contrast, the use of RAG for patient-case retrieval, where models draw on previous clinical encounters notes or EHR-derived cases, remains largely unexplored. This gap is particularly evident in the pre-diagnostic stage of GPs, where GPs must interpret sparse structured data to justify and document initial decisions.

2.2. QA Collections for Early Clinical Decision-making

Benchmark QA collections are essential for evaluating LLMs and RAG systems, particularly in domain-specific settings such as healthcare. Existing collections, such as MedQA, MedMCQA, and PubMedQA (Jin et al., 2021; Pal et al., 2022; Jin et al., 2019) typically consist of multiple-choice questions

or short patient histories designed to assess models' abilities to answer clinical questions. However, these datasets generally provide only question-response pairs, without capturing the guideline-grounded justifications that clinicians are expected to provide when formulating or documenting medical decisions. In real-world practice, especially in GP settings, transparency regarding the rationale and clinical guideline alignment behind a decision is necessary for accountability and communication with patients (Kim et al., 2024).

In an effort to move towards this direction, Kim et al. (2024) introduced MedExQA, which augments QA pairs with manually crafted explanations describing the clinical rationale behind each answer. Another example is PMC-Patients (Zhao et al., 2023), a large patient-summary retrieval dataset designed for retrieval-based clinical decision support, which supports patient-to-patient and patient-to-article retrieval tasks. This setup is conceptually similar to our approach of retrieving clinically similar patient cases to inform new decisions. While MedExQA and PMC-Patients represent important steps, their focus remains on specialist or post-diagnostic scenarios, and its explanations are generated through manual supervision rather than structured EHR data and formalized clinical guidelines.

3. TabMedQA Framework Methodology

In TabMedQA, the process of constructing question-medical encounter note pairs (QA), follows a systematic pipeline consisting of four main components: EHR Pre-processing, Question Generation, Guideline Integration, and Encounter Note Synthesis. We illustrate the overall workflow in Figure 2. To demonstrate the applicability of the framework, we apply TabMedQA to a use case focuses on prostate cancer (Saha et al. (2024)), which is described in the following subsections.

3.1. Prostate Imaging-Cancer Collection

Our framework is applied to the Prostate Imaging-Cancer collection (PI-CAI) introduced by Saha et al. (2024). PI-CAI² is a multi-modal benchmark for evaluating the detection of clinically significant prostate cancer (csPC) in a specialist diagnostic setting. The original collection³ comprises 1,500 MRI examinations from 1,476 male patients clinically suspected of having csPC and referred for specialist testing between 2012 and 2021. The

²<https://pi-cai.grand-challenge.org/>

³<https://pi-cai.grand-challenge.org/DATA/>

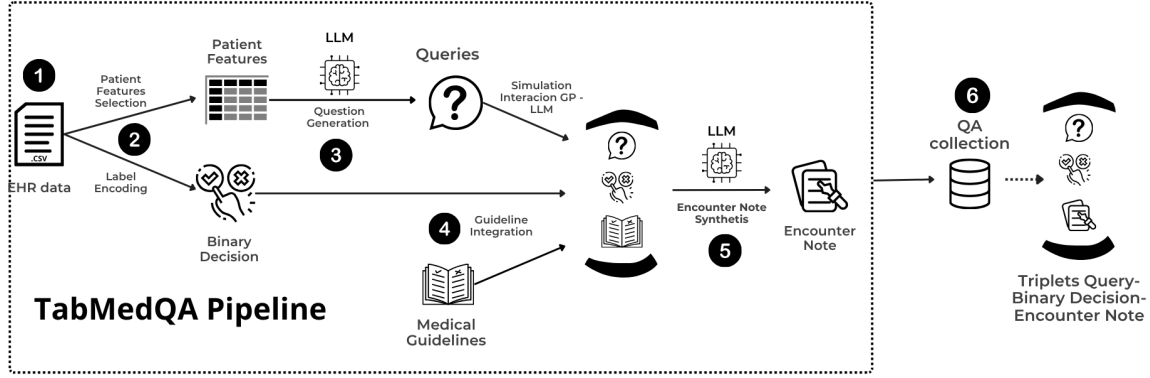


Figure 2: Overview of the **TabMedQA** pipeline: From structured (1) EHR data, (2) *EHR Preprocessing* (patient feature selection and label encoding), (3) *Question Generation*, (4–5) *Guideline Integration, Encounter Note Synthesis*, and (6) the resulting *QA collection*.

data were collected from several hospitals in the Netherlands (Saha et al., 2024; Gade et al., 2024).

Each patient case includes both imaging data and an EHR⁴ file describing clinical and demographic patient features together with a expert-verified label indicating whether follow-up was necessary under the suspicion of csPC presence (Fernandez-Quilez et al., 2024). This decision label is leveraged as the patient follow-up recommendation ground truth in our pipeline (Figure 2).

For the purpose of our work, we apply the following inclusion criteria to the original available collection (i) EHR records with no missing values for prostate specific antigen (PSA), prostate specific antigen density (PSAd), prostate volume, patient age or follow-up decision ground truth (ii) availability of all these features at patient’s first recorded consultation. After applying these criteria, the resulting collection use for our study case comprised 606 patient EHR records from different patients (Figure 2, Step 1).

3.2. EHR Preprocessing

The EHR Pre-processing stage (Figure 2, Step 2) is designed to prepare structured patient features for TabMedQA by retaining only the features that mirror the type of information GPs access during early diagnostic evaluations in primary care. Therefore, the following features are retained: patient age, PSA, PSAd, and prostate volume. The identification of these variables is conducted in collaboration with an expert and in alignment with clinical guidelines (Cornford et al., 2025).

The original PI-CAI decision label indicating whether each case is at risk of harboring csPC and thus requires follow-up, is simplified into a binary decision representing whether a patient is at risk (1) or not at risk (0) of harboring csPC.

3.3. Question Generation

The Question Generation stage (Figure 2, Step 3) aims to emulate the interaction between a GP and an LLM during an early-stage consultation, where the GP questions the system about the likelihood of a condition and its justification based on the available patient features. We choose to focus on likelihood estimation rather than binary decision based on real clinical workflows where GPs interpret these likelihoods against guideline-based risk thresholds to decide whether follow-up is necessary (Cohen et al., 2019). Each question is grounded in the patient’s structured EHR data, which provide the patient features available at the time of the consultation.

To operationalize this process, we employ a pre-trained LLM, `gpt4o_mini`⁵ to produce medically relevant questions align with the emulated GP-LLM interaction. Although the proposed methodology is model-agnostic and can be with any instructed-tuned LLM, `gpt4o_mini` is selected for its demonstrated ability to generate high-quality and contextually relevant clinical QA pairs (Lunardi et al., 2025; Keat et al., 2024; Rewthamrongsris et al., 2025).

The model is prompted with a set of structured EHR features and instructed to generate clinically coherent questions that a GP might pose to the system when evaluating a new patient and writing the respective encounter note. The Prompt 1 is designed to ensure that generated questions are (i) grounded in the provided patient features, (ii) phrased in natural language and (iii) medically plausible. All generated questions are subsequently reviewed by a clinical expert to verify their validity and relevance.

⁴https://github.com/DIAGNiJmegen/picai_labels

⁵GPT 4o mini Cost Efficient

Prompt 1: Query Generation

You are a General Practitioner expert. Given the following patient data columns:

{patient_feature_descriptions}

transform the provided tabular patient features into a medical question intended for General Practitioners and generate a JSON with the following format:

```
{"question": <generated_question>}
```

EHR data: {patient_feature_values}

Your output:

3.4. Guideline Integration and Encounter Note Synthesis

Once the question is generated, TabMedQA synthesizes complete medical encounter notes anchored in clinical guidelines, combining guideline-grounded justification with the corresponding follow-up recommendation (Figure 2, Steps 4–5). This process emulates how experts justify follow-up recommendations in real consultations.

To incorporate the guidelines, we first adapt them to the target use case in collaboration with a clinical expert (Figure 2, Step 4). This integration step serves as a connection between the structured input and the model’s ability to produce interpretable justifications, regardless of the disease or dataset being used.

For the PI-CAI study, we employ the European Association of Urology (EAU) diagnostic guidelines (Cornford et al., 2025) which define recommended threshold values for PSA and PSA_d across age groups. Due to the full EAU guidelines extend beyond what GPs typically apply during early consultations, we derive a simplified version verified by an expert capturing the practical recommendations that guide GP decision-making⁶. Whilst we refer to the study case, the approach is disease-agnostic: when official clinical guidelines are available for a condition they can be incorporated directly into the framework.

The final encounter note is then constructed (Figure 2, Step 5) by combining three elements: (i) the original GP-style question, which provides contextual understanding of what must be justified; (ii) the relevant guideline, which constrains the justification to evidence-based thresholds and recommendations; and (iii) the patient features, which anchors the justification in individual clinical findings. We employ `gpt4o_mini` as the generation model, guided by the Prompt 2.

⁶Our adapted guidelines are available at <https://anonymous.4open.science/r/1rec2026-tabmedqa-40B9>.

Prompt 2: Encounter Note Synthesis

You are a General Practitioner expert writing an encounter note for a patient. Use the provided patient features and binary decision to justify the outcome according to the guidelines: {guidelines}.

For the question "{question}", where the decision answer is {binary_decision}. Here, (1) indicates that the patient is prone to developing the disease in the future; (0) otherwise.

Provide a brief justification for General Practitioners, referencing how the patient features in the question correspond with the guidelines and considerations in clinical decision-making. Do not mention the label itself or use definitive diagnostic statements, but consider that the label indicates whether the patient has clinically significant prostate cancer or not. Use formal clinical language and relevant medical context, and do not include any information beyond what is provided.

Return your response in the following JSON format only:

```
{ "encounter_note": "<justification according to patient features and guidelines>" }
```

Your output:

The prompt results in a synthesized encounter note, consisting of a concise justification and a follow-up recommendation. These synthesized encounter notes represent the final output of TabMedQA, combining medical evidence-based justifications with follow-up recommendations to simulate a realistic GP documentation.

4. PI-CAI QA Collection

After applying TabMedQA to the PI-CAI collection, we obtain the **PI-CAI QA collection**. PI-CAI QA is composed of 606 patient records represented as triplets (q, d, n) , corresponding to the question, a binary decision, and the encounter note, respectively (Figure 2, Step 6).

To support experimental evaluation, we randomly divide the collection into a training and a test set that preserved the original preprocessed PI-CAI class distribution (1: 49%, 0: 51%). The training set consists of 484 unique triplet samples ($\approx 80\%$), and serves as a bank of past patient uses cases. The test set contains 122 unique triplet samples ($\approx 20\%$). The question q in the test set serves as the retrieval input, the corresponding encounter note n and the binary decision d act as the target ground truth. Detailed statistics of PI-CAI QA are presented in Table 1, whilst Table 2 illustrates one

randomly selected triplet from the generated PI-CAI QA collection.

5. Experiments

To verify the usability of the collection, we apply several models to it. The experiments comprise two tasks that seek to reflect the dual goals of TabMedQA: (i) binary decision classification, and (ii) generation of encounter notes that simulate the ones a GP might write after a consultation including both the decision and its corresponding clinical guideline-grounded justification.

Model Descriptions. We evaluate four models representing both feature-based and LLM-based approaches:

- **KNN:** A-nearest neighbors classifier trained on structured EHR data including patients' age, PSA, PSA_d and prostate volume. Each test case is labeled according to the majority class among k most similar training patient, computed via Minkowski distance over structured patient features. The optimal number of neighbors ($k = 19$) is selected through grid search using 10-fold cross validation.
- **KNN leveraging Encounter Note Embeddings:** Encounters notes from the training set and test questions are embedded using Clinical ModernBERT⁷ (Lee et al., 2025). We select this model due to its strong performance on clinical text embeddings, and for ensuring a fair comparison with the RAG that also relies on the same model. The classification task is performed by comparing each test question to the embedded encounter notes in this shared vector space. The best configuration is $k = 41$, also determined via grid search with 10-fold cross-validation.
- **Prompting:** The question is directly provided to `gpt-4o-mini` without retrieval. The model must infer both the likelihood of disease and its justification to synthesize the encounter note directly from EHR data.
- **RAG:** The question is augmented with the top-5 most similar encounter notes ($k = 5$) retrieved from the training set using cosine similarity over Clinical ModernBERT embeddings. The retrieved encounter notes are concatenated with the question as supporting evidence from past cases evidence (Figure 1) for generation using `gpt_4o_mini` (temperature = 0). This setup follows standard practice in

RAG-based QA configurations (Lewis et al., 2020; Asai et al., 2024)

Task 1: Binary Decision. The first task evaluates whether the models can correctly identify patients at high risk of developing csPC and requiring follow-up. Each model outputs a likelihood representing a disease risk estimation. The provided likelihood is then converted into a binary decision label using a fixed threshold and further incorporated into the generated encounter note after thresholding. This setup mirrors how a GP interprets likelihood against risk thresholds when deciding on further follow as depicted in Figure 1.

For consistency, all models produce a likelihood during inference: for KNN, the likelihood corresponds to the fraction of most similar patients with a positive binary decision; for the LLM-based approaches (Prompting and RAG), it is derived from the model's predicted likelihood or explicit numerical estimate of disease risk. For evaluation purposes, in all cases the threshold was set to 0.5.

Task 2: Encounter Note Synthesis. The second task evaluates the models' ability to synthesize encounter notes that emulate the notes a GP writes after consultation. In the Prompting scenario, the LLM synthesizes a full encounter note directly from the given question, integrating both the likelihood statement and its justification. In contrast, the encounter note synthesis in RAG is conditioned on the retrieved encounter notes from the training set, which serve as a Supporting Evidence.

Evaluation. Model performance is assessed using both automatic and manual evaluation by a clinical expert with over 5 years of clinical practice experience. The automatic evaluation consists of the binary classification task for predicting whether the patient is likely to develop csPC and thus requires follow-up. The classification is evaluated using balanced accuracy, F1, specificity, sensitivity, and AUC metrics to quantify the model's discrimination ability between patients requiring follow-up and does that do not (Fernandez-Quilez et al., 2024).

To evaluate the quality of the encounter notes and whether the supporting cases are relevant and similar to the given questions, the manual evaluation is conducted on a subset of 25 cases from the test by an clinical expert. The clinical expert reviews the retrieved supporting cases and the model-predicted encounter notes, answering a structured of Yes/No questions that assess whether the follow-up recommendation is appropriate, the likelihood estimate clinically reasonable, the justification is clear and whether the supporting retrieved cases are relevant. This evaluation complements the automatic binary decision evaluation and clinical note

⁷https://huggingface.co/Simonlee711/Clinical_ModernBERT

Statistic	Value	Description
Number of patient sessions	606	Total generated QA entries.
Number of patient features	4	Age, PSA, PSAd, prostate volume.
Binary classes	2 (0 / 1)	Non-prone vs. prone to clinically significant PCa.
Class distribution	1 / 0 (49% / 51%)	Balanced label ratio.
Training / Test split	484 / 122 (80% / 20%)	Stratified by decision label.
Avg. question length	68.24 tokens	Mean number of tokens per question.
Avg. justification length	165.1 tokens	Mean number of tokens per encounter note.
Total number of tokens	141,229	Including questions and encounter notes.
Total triplets	606	Number of (q, d, n) entries.

Table 1: Descriptive statistics of the PI-CAI QA dataset.

Field	Example
Question	For a 61-year-old male patient with a serum prostate-specific antigen level of 20.0 ng/mL, a prostate-specific antigen density of 0.52 ng/mL/mL, and a prostate volume of 38.0 mL, what is the likelihood of this patient harboring clinically significant prostate cancer?
Binary Decision	1 (patient likely to develop clinically significant prostate cancer)
Encounter Note	The PSA level exceeds the age-specific threshold of 4.5 ng/mL and the PSAd is above 0.15 ng/mL/mL, indicating a high-risk profile. A diagnostic follow-up is recommended according to the EAU guidelines.

Table 2: Example from the PI-CAI QA collection. Each triplet includes a question, the ground truth for the binary decision, and a guideline-grounded encounter note.

quality by verifying the human-based clinical plausibility and guideline alignment of the predicted notes. Each synthesized encounter note is also compared with its corresponding ground truth encounter note from the test set using BLEU and ROUGE scores, which capture surface-level textual overlap but not deeper clinical correctness or judgment quality. Results from applying these metrics are depicted in Table 5.

All experiments employ the same stratified train/test split described in Table 1, ensuring that the original dataset distribution of decision labels (i.e. 0 and 1 binary decisions) are preserved across train and test sets.

6. Results and Discussion

As shown in Table 4, the KNN baseline performs only above random (Bal. Acc. = 0.57, AUC = 0.6), suggesting that the age, PSA, PSAd and prostate volume features alone cannot clearly separate patient risk levels without additional context (e.g., most similar encounter notes). In contrast, Prompting achieves the best performance (Bal. Acc. = 0.68, AUC = 0.69), indicating that the LLM likely leverages medical information captured during pre-training. RAG reaches similar balanced accuracy but shows higher sensitivity (0.87) and lower specificity (0.29) implying that retrieval increases the model’s ability to detect positive binary decisions but at the cost of introducing more false positives. Although the scores are modest, they are comparable to art results on this collection and

reflect the difficulty of early-stage PCa follow-up decision-making (Fernandez-Quilez et al., 2024).

Table 3 shows examples of predicted encounter notes synthesized with Prompting and RAG settings to compare their generated clinical justifications and follow-up recommendations. During the manual evaluation by the clinical expert, all follow-up recommendations (% 100) in the tested subset are judged as clinically appropriate by the expert, meaning that the model’s advice for further tests or monitoring matched with PSA and PSAd thresholds in the guidelines. On the other hand, the estimated likelihood suggested by the system is deemed correct in only about half of the cases (% 52). Around half of the justifications (% 52) are also considered clinically clear and aligned with clinical expectations, while most retrieved supporting cases (i.e. similar encounter notes) are assessed as relevant and helpful for synthesizing new patient encounter notes (% 92).

When looking at the retrieved encounter notes in more detail, 100 out of 125 individual retrieved cases (i.e. 5 cases retrieved for each of 25 manually evaluated notes) are marked as useful. After close inspection, we observe that the most of the incorrectly retrieved cases contain big differences between retrieved PSA and PSAd values when compared to those of the patient under consideration. These differences reveal models’ sensitivity to small variations in patient features, which must be further analyzed to understand what patterns the model relies on and how retrieval affects its judgments. Overall, the results show that even though

Method	Example
Prompting	The patient's age and elevated PSA level suggest a higher risk for prostate cancer. A PSA of 9.3 ng/mL is above the typical threshold for concern, and while the PSA _d of 0.25 ng/mL/mL is moderate, the prostate volume is normal. These factors indicate a significant possibility of clinically relevant prostate cancer, but further diagnostic evaluation is needed for a definitive determination.
RAG	The patient is 69 years old with a PSA level of 9.3 ng/mL and a PSA _d of 0.25 ng/mL/mL, both of which are elevated. Comparing to Case 5 (69 years, PSA 17.8 ng/mL, PSA _d 0.23 ng/mL/mL) and Case 1 (PSA 9.4 ng/mL, PSA _d 0.25 ng/mL/mL), the patient has similar elevated PSA and PSA _d values, which suggest a higher risk for clinically significant prostate cancer. The prostate volume of 37.0 mL is also within the range of concern, supporting the need for further assessment of significant pathology.

Table 3: Examples of encounter notes generated by Prompting and RAG models at the inference stage.

Method	Bal. Acc.	F1	Spec.	Sens.	AUC
KNN	0.57	0.58	0.54	0.60	0.63
KNN + Encounter Notes Embeddings	0.57	0.56	0.58	0.55	0.56
Prompting	0.68	0.69	0.63	0.73	0.69
RAG	0.58	0.67	0.29	0.87	0.68

Table 4: Comparison of the models on the binary classification of csPA on the PI-CAI QA dataset. Metrics reported include Balanced Accuracy (Bal. Acc.), F1-score (F1), Specificity (Spec.), Sensitivity (Sens.), and Area Under the ROC Curve (AUC).

Method	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Prompting	0.09	0.50	0.26	0.34
RAG	0.12	0.58	0.31	0.39

Table 5: Comparison of explanations and ground truth n-gram overlaps based on BLEU and ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) scores for the LLM-based methods.

the reliability of the encounter notes is still limited, the RAG is able to quite reliably provide the supporting cases, what might be crucial for subsequent GP’s recommendation.

7. Conclusion and Future Work

In this work, we introduce **TabMedQA**, a framework for constructing medical QA collections directly from structured EHRs. By combining instruction-tuned LLMs with disease-specific clinical guidelines, the framework enables the automated generation of clinical encounter notes that simulate how GPs formulate and document early-stage clinical decisions, including the clinically-grounded justifications that support them.

We demonstrate the applicability of **TabMedQA** in a prostate cancer case study, where the synthesized encounter notes include both guideline-based justifications and follow-up recommendations. Experiments show that LLM-based methods make more clinically accurate decisions than feature-based baselines, even though the justifications are not fully clinically correct, factual, or clear. On the other hand, manual expert evaluations show that RAG can reliably retrieve supporting cases and thus enables more reliable human machine-aided

decision making.

Future work will extend the framework to additional diseases and multi-modal EHR data, explore automatic factuality and guideline-adherence metrics, and investigate temporally aware RAG setups where retrieved clinical evidence evolves over time.

8. Limitations

Despite of the effectiveness of our TabMedQA framework, certain limitations still remain. First, since the questions are derived from structured EHRs rather than real GP-patient interactions, the linguistic diversity of the generated text remain limited. Even though an LLM is employed to enrich phrasing and emulate human questions. As a consequence, the resulting language may still reflect the structured nature of EHRs rather than genuine GP questions. In the same direction, the quality and coherence of the synthesized encounter notes are also dependent on the specific LLM being used during the collection construction, which may introduce model biases or inconsistencies. While our framework leverages medical guidelines, its reliability still depends on expert oversight to ensure that the synthesized justification and follow-up recommendations are truly guidelines-consistent.

Finally, TabMedQA is currently demonstrated on a single structured collection (PI-CAI), which limits its generalization across other clinical domains and data sources.

9. Ethics Statements

This research study was conducted using retrospectively open-access human subject data (Saha et al., 2024). The use was approved by the institutional review board of all PI-CAI centers.

10. Bibliographical References

- Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877.
- Akari Asai, Zequi Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.
- 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:e53636.
- Genna R Cohen, Charles P Friedman, Andrew M Ryan, Caroline R Richardson, and Julia Adler-Milstein. 2019. Variation in physicians' electronic health record documentation and potential patient harm from that variation. *Journal of general internal medicine*, 34(11):2355–2367.
- P. Cornford, D. Tilki, R. C. N. van den Bergh, et al. 2025. Diagnostic evaluation. in *EAU-EANM-ESTRO-ESUR-ISUP-SIOG Guidelines on Prostate Cancer*. European Association of Urology Guidelines. Limited update March 2025. Available at: <https://uroweb.org/guidelines/prostate-cancer/chapter/diagnostic-evaluation>, accessed September 8, 2025.
- Jesse C Crosson, Christine Stroebel, John G Scott, Brian Stello, and Benjamin F Crabtree. 2005. Implementing an electronic medical record in a family medicine practice: communication, decision making, and conflict. *The Annals of Family Medicine*, 3(4):307–311.
- Hans Eguia, Carlos Luis Sánchez-Bocanegra, Franco Vinciarelli, Fernando Alvarez-Lopez, and Francesc Saigó-Rubió. 2024. Clinical decision support and natural language processing in medicine: systematic literature review. *Journal of Medical Internet Research*, 26:e55315.
- Ran Elgedawy, Ioana Danciu, Maria Mahbub, and Sudarshan Srinivasan. 2024. Dynamic q&a of clinical documents with large language models. *arXiv preprint arXiv:2401.10733*.
- Wonca Europe. 2002. The european definition of general practice/family medicine. *Barcelona: WONCA Europe*.
- Alvaro Fernandez-Quilez, Tobias Nordström, Fredrik Jäderling, Svein Reidar Kjosavik, and Martin Eklund. 2024. Prostate age gap: an mri surrogate marker of aging for prostate cancer detection. *Journal of Magnetic Resonance Imaging*, 60(2):458–468.
- Farieda Gaber, Maqsood Shaik, Fabio Allega, Agnes Julia Bilecz, Felix Busch, Kelsey Goon, Vedran Franke, and Altuna Akalin. 2025. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8(1):263.
- Marius Gade, Kevin Mekhaphan Nguyen, Sol Gedde, and Alvaro Fernandez-Quilez. 2024. Impact of uncertainty quantification through conformal prediction on volume assessment from deep learning-based mri prostate segmentation. *Insights into Imaging*, 15(1):286.
- Jin Ge, Steve Sun, Joseph Owens, Victor Galvez, Oksana Gologorskaya, Jennifer C Lai, Mark J Pletcher, and Ki Lai. 2024. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. *Hepatology*, 80(5):1158–1168.
- Cesar A Gomez-Cabello, Sahar Borna, Sophia Pressman, Syed Ali Haider, Clifton R Haider, and Antonio J Forte. 2024. Artificial-intelligence-based clinical decision support systems in primary care: A scoping review of current clinical implementations. *European Journal of Investigation in Health, Psychology and Education*, 14(3):685–698.
- 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. *arXiv preprint arXiv:1909.06146*.

- Yu He Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Chang-Fu Kuo, et al. 2025. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine*, 8(1):187.
- Karl Keat, Rasika Venkatesh, Yidi Huang, Rachit Kumar, Sony Tuteja, Katrin Sangkuhl, Binglan Li, Li Gong, Michelle Whirl-Carrillo, Teri E Klein, et al. 2024. Pgxqa: A resource for evaluating llm performance for pharmacogenomic qa tasks. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 229–246. World Scientific.
- Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. Medexqa: Medical question answering benchmark with multiple explanations. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 167–181.
- Felipe C Kitamura, Luciano M Prevedello, Errol Colak, Safwan S Halabi, Matthew P Lungren, Robyn L Ball, Jayashree Kalpathy-Cramer, Charles E Kahn Jr, Tyler Richards, Jason F Talbott, et al. 2024. Lessons learned in building expertly annotated multi-institution datasets and hosting the rsna ai challenges. *Radiology: Artificial Intelligence*, 6(3):e230227.
- Robert Korom, Sarah Kiptinness, Najib Adan, Kassim Said, Catherine Ithuli, Oliver Rotich, Boniface Kimani, Irene King'ori, Stellah Kamau, Elizabeth Atemba, et al. 2025. Ai-based clinical decision support for primary care: A real-world study. *arXiv preprint arXiv:2507.16947*.
- Simon A Lee, Anthony Wu, and Jeffrey N Chiang. 2025. Clinical modernbert: An efficient and long context encoder for biomedical text. *arXiv preprint arXiv:2504.03964*.
- 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.
- Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. 2024. Clinicalrag: enhancing clinical decision support through heterogeneous knowledge retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 64–68.
- Riccardo Lunardi, Michael Soprano, Paolo Coppola, Vincenzo Della Mea, Stefano Mizzaro, and Kevin Roitero. 2025. Pils of knowledge: A synthetic benchmark for evaluating question answering systems in healthcare. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3648–3658.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.
- Stephan Rau, Alexander Rau, Johanna Nattenmüller, Anna Fink, Fabian Bamberg, Marco Reiser, and Maximilian F Russe. 2024. A retrieval-augmented chatbot based on gpt-4 provides appropriate differential diagnosis in gastrointestinal radiology: a proof of concept study. *European radiology experimental*, 8(1):60.
- Paak Rewthamrongsris, Jirayu Burapachee, Vorapat Trachoo, and Thantrira Porntaveetus. 2025. Accuracy of large language models for infective endocarditis prophylaxis in dental procedures. *International Dental Journal*, 75(1):206–212.
- Deborah Rose and Nikita Joshi. 2018. The digital doctor: Hope, hype, and harm at the dawn of medicine's computer age. *Annals of Emergency Medicine*, 71(1):154–155.
- Anindo Saha, Joeran S Bosma, Jasper J Twilt, Bram van Ginneken, Anders Bjartell, Anwar R Padhani, David Bonekamp, Geert Villeirs, Georg Salomon, Gianluca Giannarini, et al. 2024. Artificial intelligence and radiologists in prostate cancer detection on mri (pi-cai): an international, paired, non-inferiority, confirmatory study. *The Lancet Oncology*, 25(7):879–887.
- Sina Shool, Sara Adimi, Reza Saboori Amleshi, Ehsan Bitaraf, Reza Golpira, and Mahmood Tara. 2025. A systematic review of large language model (llm) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1):117.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.
- Dean F Sittig and Hardeep Singh. 2011. Legal, ethical, and financial dilemmas in electronic health record adoption and use. *Pediatrics*, 127(4):e1042–e1047.

Barbara Starfield, Leiyu Shi, and James Macinko. 2005. [Contribution of primary care to health systems and health](#). *The Milbank Quarterly*, 83(3):457–502.

Steven M Williamson and Victor Prybutok. 2024. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in ai-driven healthcare. *Applied Sciences*, 14(2):675.

Gongbo Zhang, Zihan Xu, Qiao Jin, Fangyi Chen, Yilu Fang, Yi Liu, Justin F Rousseau, Ziyang Xu, Zhiyong Lu, Chunhua Weng, et al. 2025. Leveraging long context in retrieval augmented language models for medical question answering. *npj Digital Medicine*, 8(1):239.

Jiacheng Zhao, Zihan Li, Ziqi Wang, Yan Song, and Yue Zhang. 2023. [Pmc-patients: a large-scale dataset of patient summaries and relations for retrieval-based clinical decision support](#). *Scientific Data*, 10(1):814.

SimpliMED: Automatic Simplification of Cardiology Discharge Reports Using Large Language Models

Lucas Molino Piñar¹, Manuel Carlos Díaz Galiano¹,
María Teresa Martín Valdivia¹, José Ángel Urbano-Moral², Elena Sola-García²

¹SINAI Research Group, University of Jaén, Spain

²ReCaH Research Group, University Hospital of Jaén, Spain

¹{lmolino, mcdiaz, maite}@ujaen.es, ²{jaurbanomoral, esola.cardiologia}@gmail.com

Abstract

Medical discharge reports frequently contain highly technical language that creates significant communication barriers between healthcare professionals and patients, potentially compromising treatment adherence and post-discharge care quality. In this paper, we present SimpliMED, a modular system designed to automatically simplify cardiology discharge reports using Large Language Models (LLMs) and advanced Natural Language Processing techniques (NLP). Our architecture integrates section-based preprocessing with specialized prompts, explicit handling of medical abbreviations, and therapeutic explanations of medications to enhance accessibility. We evaluate our system using a corpus of 307 anonymized cardiology discharge reports from a Spanish medical center. For abbreviation detection, our fine-tuned Small Language Model (SLM) achieves an F1-score of 0.90, significantly outperforming regex-based approaches (F1: 0.67). For medication recognition, we achieve F1-scores of 0.91 for commercial names and 0.70 for active principles. We also contribute a therapeutic dictionary containing 14,611 medications with patient-friendly explanations extracted from the Spanish Agency of Medicines. Expert evaluation by two cardiologists yields an overall quality score of 75%, with highest performance for admission reason (91%) and current illness (75%) sections. While results demonstrate the potential of LLM-based medical text simplification for Spanish clinical language, we identify areas requiring further development before clinical deployment.

Keywords: Medical text simplification, clinical NLP, large language models, discharge reports, patient accessibility, Spanish clinical language, plain language

1. Introduction

Effective medical communication is fundamental to the relationship between healthcare professionals and patients. However, hospital discharge reports frequently employ specialized technical language that hinders comprehension by patients and their families (Kandula et al., 2010). This communication barrier not only generates uncertainty but can also compromise treatment adherence and subsequent medical follow-up. When patients leave the hospital without fully understanding their diagnosis, prescribed treatments, or warning signs requiring immediate attention, the consequences can range from medication errors to delayed recognition of serious complications.

This challenge is particularly relevant given that a significant portion of the population has limited reading competence (OECD, 2023). The demand for accessible texts has led to initiatives such as Easy-to-Read guidelines (AENOR, 2018) and the Plain Language movement (Adler, 2012). Among the first systems addressing text simplification for Spanish, Bott and Saggion (2012) proposed an automatic approach targeting e-Accessibility. Recently, the Spanish Organic Law 5/2024 on the Right of Defense mandates that communications be drafted in clear, universally accessible language, highlighting the growing legal recognition of text ac-

cessibility.

Recent advances in NLP, particularly the development of LLMs (Brown et al., 2020), open new possibilities for addressing this challenge. These tools are capable of transforming complex texts while maintaining their precise content and adapting the register and complexity of language according to the receiver's needs (Joseph et al., 2023). The emergence of the Transformer architecture (Vaswani et al., 2017) revolutionized the field, enabling models that capture long-range dependencies and generate coherent, contextually appropriate text.

In this paper, we present SimpliMED, an automated system designed to transform cardiology discharge reports into versions that are accessible and comprehensible to patients. Cardiology reports present particular challenges: dense technical terminology, numerous abbreviations with context-dependent meanings, complex medication regimens, and temporal narratives describing clinical evolution. Through advanced NLP techniques, we seek to develop a solution that preserves the integrity of medical information while facilitating its interpretation.

Our main contributions include a modular architecture for medical report simplification combining LLMs with specialized modules for abbreviation and medication handling (Figure 1), featuring section-

based preprocessing that enables tailored simplification for each report component. We developed two fine-tuned Small Language Models (SLMs) achieving F1-scores of 0.90 for clinical abbreviation detection and 0.91 for medication recognition, as well as a therapeutic dictionary containing 14,611 medications with simplified explanations extracted from the AEMPS (*Agencia Española de Medicamentos y Productos Sanitarios*—Spanish Agency of Medicines and Medical Devices)¹. Finally, we present expert clinical validation by cardiologists with structured evaluation methodology.

The remainder of this paper is organized as follows: Section 2 reviews related work in text simplification. In section 3 we describe the SimpliMED system architecture. Section 4 presents the experimental setup, showing the results in Section 5. Section 6 discusses limitations. Finally we conclude with Section 7.

2. Related Work

The field of automatic text simplification has undergone a paradigm shift in recent years, evolving from early rule-based systems to the current dominance of Large Language Models (LLMs). As detailed in the systematic review by Bednarczyk et al. (2025), this evolution reflects a transition from rigid syntactic operations towards semantic-aware generative approaches. Their analysis of 45 recent studies concludes that LLMs significantly outperform previous architectures in clinical summarization and simplification tasks, particularly when fine-tuned on domain-specific corpora.

Historically, simplification relied on manual rules and synonym replacement (Carroll et al., 1998), an approach that, while interpretability-friendly, struggled with the rich variability of natural language. The statistical machine translation era (Wubben et al., 2012) and subsequent neural sequence-to-sequence models (e.g., T5) (Botarleanu et al., 2020) improved fluency but often required massive parallel corpora, Simplification/Original pairs, which are notoriously scarce in specialized domains like medicine.

In the healthcare context, simplification is not merely a readability task but a patient safety necessity. Kandula et al. (2010) identified early on that medical simplification requires handling not just vocabulary but also syntactic complexity and logical explanations. Unlike general domain simplification, medical applications must prioritize content preservation above all else. Bednarczyk et al. (2025) emphasize that while general-purpose models (like GPT-4 or LLaMA) exhibit strong zero-shot capabilities, superior performance in clinical settings is

often achieved through specialized fine-tuning. The comprehensive survey by Ondov et al. (2022) further systematizes automated methods for biomedical text simplification, identifying key challenges including terminology handling, evaluation methodology, and the scarcity of domain-specific parallel corpora.

Recent work addresses simplification from complementary perspectives. Rust et al. (2025) demonstrated that LLMs can effectively simplify discharge summaries while augmenting them with lifestyle recommendations. Vázquez-Rodríguez et al. (2023) proposed document-level simplification with coherence evaluation, highlighting the importance of maintaining discourse structure. Vázquez-Rodríguez et al. (2021) further investigated evaluation methodologies for text simplification, revealing limitations of standard metrics. Notably, Xia et al. (2025) introduced JEBS, a fine-grained biomedical lexical simplification task that explicitly separates the identification of hard-to-understand terms from the simplification process, an approach conceptually aligned with our modular architecture. However, Joseph et al. (2023) warn that these models require rigorous domain-specific evaluation beyond standard metrics like BLEU (Papineni et al., 2002) or SARI (Xu et al., 2016), which correlate poorly with clinical accuracy.

Despite these global advances, resources for Spanish medical simplification remain limited compared to English. The CLARA-MeD project (Campillos-Llanos et al., 2022) represents a significant milestone, providing the first comparable corpus for Spanish medical text simplification, yet the ecosystem lacks the robust pipeline tools available for English. Most existing Spanish systems rely on direct translation of English models or general-purpose LLMs without clinical grounding. Our work addresses this gap by proposing a modular architecture that explicitly separates abbreviation and medication handling from the simplification generation, utilizing the generative power of modern LLMs while imposing the structural constraints necessary for clinical reliability.

3. Methodology

This section presents the SimpliMED system architecture, including model selection rationale, the specialized modules for abbreviation and medication handling, and the prompt engineering strategy.

3.1. System Architecture

SimpliMED follows a modular pipeline architecture (Figure 1) explicitly designed to address the linguistic challenges of clinical documentation through four integrated components: preprocessing and

¹<https://cima.aemps.es/>

segmentation, abbreviation detection, medication recognition, and LLM-based simplification.

Our approach prioritizes the structural modularity of clinical reports, recognizing that each section serves a distinct communicative function (Denny et al., 2009). For instance, the admission reason typically introduces the clinical episode concisely, whereas diagnostic tests present technical results requiring detailed explanation. Consequently, the system processes each section independently using specialized prompts (Figure 2). This granular processing strategy allows the incorporation of appropriate levels of detail, structural formatting, and explanatory depth specifically suited to the unique requirements of each section type, ensuring higher quality simplification compared to global processing approaches.

To further enhance accessibility, the architecture incorporates specialized mechanisms for handling technical terminology. Medical reports frequently employ abbreviations like AF (Atrial Fibrillation—abbreviated as “FA” in Spanish), HT (HyperTension—abbreviated as “HTA” in Spanish), and LVEF (Left Ventricular Ejection Fraction—abbreviated as “FEVI” in Spanish) which create significant barriers for patients. The system addresses this issue by explicitly detecting and expanding these terms, maintaining the original abbreviation in parentheses to facilitate learning while ensuring comprehension. Similarly, treatment sections often present medications in condensed formats with cryptic dosing schedules. The system enriches this information by identifying medications and retrieving their specific therapeutic purposes (e.g., “anticoagulation”, “lipid control”), thereby providing patients with essential context about their treatment regimen.

The implementation of this pipeline begins with preprocessing and segmentation, where raw reports, often containing digitization artifacts, are transformed into structured text. This stage handles irregularities and divides the content into up to ten standard clinical sections including admission reason, medical history, current illness, events during admission, diagnostic tests, surgical interventions, discharge diagnosis, treatment, discharge plan, and follow-up. Following segmentation, the text flows through the abbreviation and medication modules (detailed in Section 3.4 and Section 3.5 respectively) before the final LLM-based simplification phase (Section 3.6).

3.2. Model Selection

After preliminary evaluation of multiple architectures including *LLaMA 2/3* variants (1B–8B parameters), *Mistral 7B*, *Gemma 7B*, *Phi-2*, and OpenAI *GPT* models (3.5/4/4o), we selected *LLaMA 3.1 8B Instruct* as our base model. The selection was

driven primarily by privacy requirements: sensitive medical data requires local processing without external API calls, and healthcare data protection regulations prohibit transmission of patient information to third-party cloud services. Consequently, non-open-weight models were discarded for production use. While OpenAI models were explored in preliminary phases for methodology development, they were not used for processing real clinical reports.

The model demonstrated exceptional contextual sensitivity, maintaining coherence across extended reports and preserving relationships between sections, crucial when findings in one section depend on information presented elsewhere, such as when the discharge diagnosis must be understood in the context of diagnostic test results. *LLaMA 3.1 8B* also showed basic clinical reasoning capacity sufficient to identify causal relationships between symptoms, findings, and recommendations, which proved adequate for the simplification task that primarily requires reformulation rather than diagnosis. Finally, the 8B parameter model offers an optimal balance between capability and computational resources for clinical deployment scenarios, with inference performed on a single GPU within acceptable time constraints.

3.3. Prompt Engineering Strategy

The design of prompts for each section follows a systematic architecture where the model assumes the role of an expert in simplifying cardiology reports for elderly patients with limited medical training. Each prompt specifies simplification rather than summarization, maintaining semantic integrity while ensuring comprehensibility. The prompts emphasize concision through short sentences, avoid unnecessary jargon, and require a formal third-person style with respectful tone.²

A key strategy involves preserving medical terms in parentheses when simplifying technical vocabulary, allowing patients to recognize original terminology. Semantic fidelity remains the main criterion: any reformulation must preserve clinical meaning. The framework provides homogeneous structure while allowing section-specific nuances, for admission reason or medical history, causal clarity is reinforced; for diagnostic tests or treatment, the need to explain technical terms is emphasized.

3.4. Abbreviation Detection Module

For abbreviation detection, we compared two approaches. The first is a regex-based detection approach using the pattern `(?<![a-zA-Z0-`

²The complete code including all section-specific prompts is available at <https://github.com/lmolino03/simplified>.

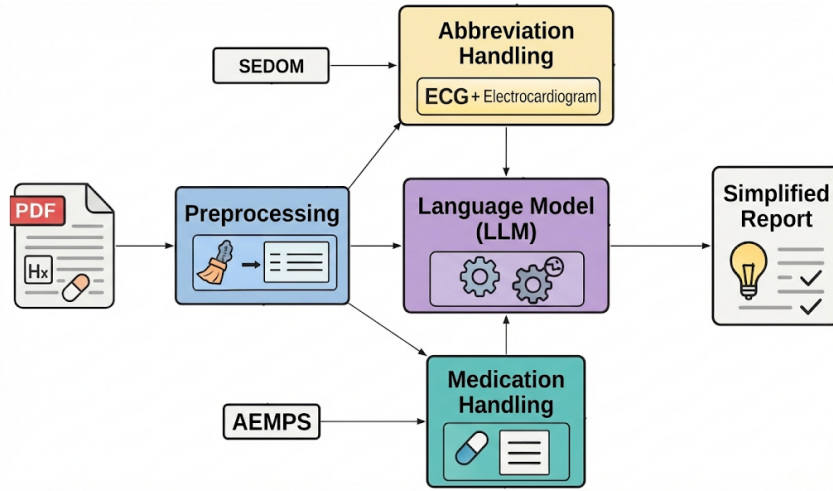


Figure 1: SimpliMED system architecture. Clinical reports are preprocessed, segmented into sections, enriched with abbreviation meanings and medication information, then simplified using specialized prompts per section type.

MOTIVO DE INGRESO
Dolor torácico.
ANTECEDENTES
HTA en tratamiento. DM tipo 2. Dislipemia. Exfumador desde hace 5 años.
ENFERMEDAD ACTUAL
Varón de 68 años que acude por dolor torácico típico con cortejo vegetativo.
EVOLUCIÓN Y CURSO CLÍNICO
Paciente que ingresa en UCI coronaria. Se realiza coronariografía urgente con implante de stent en DA proximal.
PRUEBAS COMPLEMENTARIAS
ECG: elevación ST en V1-V4. Troponinas elevadas. Ecocardiograma: FEVI 45%.
INTERVENCIÓN - PROCEDIMIENTOS QUIRÚRGICOS
Coronariografía con implante de stent farmacoactivo en DA proximal.
JUICIO CLÍNICO
SCACEST anterior. Enfermedad coronaria de un vaso (DA proximal).
TRATAMIENTO
AAS 100mg 1-0-0, Clopidogrel 75mg 1-0-0, Atorvastatina 80mg 0-0-1.
PLAN DE ACTUACIÓN
Rehabilitación cardíaca. Control de factores de riesgo cardiovascular.
REVISIONES
Consulta de Cardiología en 4 semanas. Control analítico previo.

(a) Original Report (Spanish)

REASON FOR ADMISSION
Chest pain.
MEDICAL HISTORY
HTA under treatment. Type 2 DM. Dyslipidemia. Ex-smoker for 5 years.
CURRENT ILLNESS
68-year-old male presenting with typical chest pain with autonomic symptoms.
EVENTS DURING ADMISSION
Patient admitted to coronary ICU. Urgent coronary angiography performed with stent implantation in proximal LAD.
DIAGNOSTIC TESTS
ECG: ST elevation in V1-V4. Elevated troponins. Echocardiogram: LVEF 45%.
INTERVENTION - SURGICAL PROCEDURES
Coronary angiography with drug-eluting stent implantation in proximal LAD.
DISCHARGE DIAGNOSIS
Anterior STEMI. Single-vessel coronary disease (proximal LAD).
TREATMENT
ASA 100mg 1-0-0, Clopidogrel 75mg 1-0-0, Atorvastatin 80mg 0-0-1.
DISCHARGE PLAN
Cardiac rehabilitation. Cardiovascular risk factor control.
FOLLOW-UP
Cardiology consultation in 4 weeks. Prior bloodwork.

(b) Translated Report for legibility reason(English)

Figure 2: Example of a cardiology discharge report structure showing typical sections processed by SimpliMED.

9)) ([A-Z]{2,}) (?![a-zA-Z0-9]) to identify sequences of two or more uppercase characters not adjacent to alphanumeric, with detected candidates validated against the SEDOM (*Sociedad Española de Documentación Médica*—Spanish Society of Medical Documentation) dictionary³ containing 7,054 medical abbreviation entries. The second approach uses a *LLaMA 3.2 3B Instruct* model fine-tuned on a *GPT-4o*-annotated corpus, where the training corpus was generated from 70% of the clinical reports with *GPT-4o* identifying all abbreviations present. Fine-tuning configuration included batch size 2, 3 epochs, learning rate 1×10^{-4} , max-

imum sequence length 2,048 tokens, and gradient accumulation steps 4.

The SLM approach offers advantages over regex because it considers semantic context rather than purely orthographic patterns. For example, it can distinguish “DE” as “*Desviación Estándar*” (Standard Deviation) in a laboratory results context versus the Spanish preposition “*de*” (of/from) appearing in uppercase text. For abbreviations with multiple possible meanings, we delegate disambiguation to the simplification LLM, where the prompt includes the detected abbreviation along with its possible meanings from the SEDOM dictionary, requesting the model to select the contextually appropriate interpretation based on the full report content.

³<https://sedom.es/>

3.5. Medication Recognition Module

Medication recognition relies exclusively on a fine-tuned SLM, deliberately excluding regex-based approaches. This design decision stems from the high variability in how medications appear in clinical text: misspellings, informal abbreviations, and non-standard forms are common in handwritten or rapidly dictated reports.

The *LLaMA 3B* model was fine-tuned for medication and active principle recognition using the same training configuration as the abbreviation module. The annotated corpus was generated using *GPT-4o* to identify all medication mentions in the training set.

To provide meaningful explanations of medication purposes, we constructed a therapeutic dictionary from the AEMPS database. We collected 14,611 authorized medications in Spain, extracting information from both patient leaflets and technical specifications. For each medication, *GPT-4o mini* extracted the therapeutic use from Section 1 of patient leaflets (Accessible language) and Section 4.1 of technical specifications (Clinical indications), resulting in a dictionary that links each medication and active principle to a concise therapeutic explanation. This dictionary is publicly available⁴.

Medication lookup uses Levenshtein distance for fuzzy matching to handle spelling variations and typos common in clinical text. For combined medications with multiple active principles, the system searches for entries containing all detected components.

3.6. LLM-Based Simplification Module

The final simplification stage integrates the outputs of the abbreviation and medication modules into a structured prompt for *LLaMA 3.1 8B Instruct*. For each section, the prompt includes: (1) the original clinical text, (2) a list of detected abbreviations with their expanded meanings as provided by the abbreviation module, and (3) for treatment sections, the identified medications with their therapeutic purposes retrieved from the AEMPS dictionary.

The model processes each section independently using section-specific prompts. In the “with module” configuration, the LLM receives pre-processed abbreviation expansions and medication information, enabling it to produce accurate simplifications. In contrast, the “without module” baseline relies solely on the LLM’s internal knowledge to interpret abbreviations and medications during simplification, without any prior detection or dictionary lookup step.

⁴https://github.com/sinai-uja/DICCIONARIO_MEDICAMENTOS

4. Experimental Setup

4.1. Corpus Description

Our corpus comprises 307 anonymized cardiology discharge reports from University Hospital Jaen (Spain), provided under an institutional collaboration framework and anonymized following data protection protocols. This cardiology-focused dataset presents thematic and terminological coherence that reduces linguistic variability, enabling development of focused systems that are easier to control and evaluate.

Reports follow a standard structure with up to ten sections (Table 1), though not all sections appear in every report. Sections are explicitly labeled in the reports using recognizable headers, which our preprocessing module detects through pattern matching. The most frequent sections are current illness, medical history, and diagnostic tests, appearing in approximately 95% of reports, while surgical intervention sections appear only in 42 of 307 reports (14%).

Section	Freq.	Avg. Tokens
Admission reason	98%	~5
Medical history	95%	~85
Current illness	92%	~70
Discharge diagnosis	90%	~45
Treatment	88%	~120
Diagnostic tests	75%	~180
Events during admission	70%	~150
Discharge plan	65%	~50
Follow-up	45%	~25
Surgical intervention	14%	~80

Table 1: Report section frequency and average token count.

Section length varies considerably by type. Admission reason sections are typically brief (median ~5 tokens), providing concise statements of the reason for hospitalization. In contrast, diagnostic test sections show the highest average length (~180 tokens), reflecting the need to document multiple diagnostic study results including numerical values, reference ranges, and detailed technical descriptions.

4.2. Abbreviation Analysis

Analysis of the corpus reveals systematic patterns in abbreviation usage. The most frequent abbreviations include cardiovascular-specific terms such as AF, HT, and LVEF (abbreviated as “FA, HTA, and FEVI” in Spanish).

The abbreviation density varies considerably by section type. Treatment sections show the highest abbreviation density, with an average of 8.3 abbre-

viations per 100 tokens, reflecting the condensed nature of medication prescriptions. In contrast, admission reason sections show lower density (2.1 per 100 tokens), as they typically use more complete clinical descriptions. This pattern suggests that sections requiring quick clinical reference tend to use more abbreviated forms, while patient-facing sections, those most likely to be read and interpreted by patients, such as admission reason, discharge plan, and follow-up, maintain more explicit terminology.

4.3. Evaluation Methodology

Given the absence of established automatic metrics for medical text simplification, we employ a combination of automatic evaluation for subtasks and expert medical evaluation for overall simplification quality.

For automatic evaluation of subtasks, we computed Precision, Recall, and F1-score for abbreviation detection and medication recognition. The corpus was split into 70% for training (215 reports) and 30% for testing (92 reports), with the test set held out entirely from fine-tuning. The test set was annotated using *GPT-4o*, which was prompted to identify all abbreviations and medication mentions in each report. These annotations were then manually reviewed by the first author to correct errors, particularly false positives from non-medical uppercase acronyms and missed active principles. While we acknowledge that using LLM-generated annotation, even after manual revision, introduces potential bias when evaluating other language models, this approach was necessitated by the lack of expert-annotated clinical NLP resources for Spanish, and the cost of full manual annotation by clinicians.

For expert clinical evaluation, two cardiologists evaluated simplified versions of 8 randomly selected reports. Evaluation was performed by section using a 3-level scale where 0 indicates incorrect simplifications containing factual errors or misinterpretations that could mislead patients, 0.5 indicates partial success where the simplification is generally acceptable but contains minor issues or room for improvement, and 1 indicates correct simplifications that accurately convey the original information in accessible language suitable for patients. Additionally, experts provided qualitative feedback including observations, improvement suggestions, and general comments on simplification quality. This dual approach captures both structured ratings and nuanced clinical perspectives that numerical scores cannot fully represent.

5. Results

This section presents quantitative results for the specialized subtasks (abbreviation detection, medication recognition) and qualitative results from expert evaluation of the complete simplification system.

5.1. Abbreviation Detection

The fine-tuned SLM significantly outperforms the regex-based approach (Table 2).

System	Precision	Recall	F1
Regex-based	0.63	0.77	0.67
SLM (fine-tuned)	0.89	0.93	0.90

Table 2: Abbreviation detection performance comparison.

The SLM achieves 23 percentage points higher F1-score than the regex approach, which suffers from lower precision (0.63) due to false positives from uppercase text segments that are not abbreviations.

5.2. Medication Recognition

Table 3 presents results for medication and active principle recognition.

Category	Precision	Recall	F1
Active substances	0.70	0.71	0.70
Brand names	0.91	0.91	0.91
Combined (union)	0.89	0.93	0.90

Table 3: Medication recognition performance by category.

Medication brand names achieve the highest F1 (0.91), while active substances show lower performance (0.70) due to category confusion. The combined evaluation achieves F1 of 0.90, demonstrating strong overall medication recognition.

5.3. Expert Evaluation

Two cardiologists independently evaluated simplified versions of 8 randomly selected reports, scoring each section on a 0–1 scale where 0 indicates incorrect simplifications with factual errors, 0.5 indicates partial success with minor issues, and 1 indicates correct simplifications suitable for patients. Table 4 summarizes the evaluation scores across all sections.

Performance varies considerably across section types. Admission reason achieves the highest score (0.91), reflecting brief, factual content requiring minimal interpretation, while diagnostic tests

Section	Score
Admission reason	0.91
Current illness	0.75
Discharge plan	0.71
Medical history	0.69
Intervention	0.67
Discharge diagnosis	0.56
Follow-up	0.56
Treatment	0.53
Events during admission	0.50
Diagnostic tests	0.44
Global Score	0.75

Table 4: Expert evaluation scores by section (n=8 reports, scale 0-1).

receives the lowest (0.44), as technical laboratory values and electrocardiogram readings require clinical interpretation beyond current model capabilities. The global score of 0.75 indicates reasonable quality, though significant improvement remains necessary before clinical deployment.

Analysis of inter-rater agreement yields a weighted Cohen’s Kappa of 0.72 (Landis and Koch, 1977), indicating substantial agreement, with 73% of ratings achieving exact agreement. The results suggest the model excels at reformulating declarative statements but struggles with content requiring domain expertise for interpretation.

5.4. Qualitative Examples

The following examples illustrate how each specialized module contributes to improved simplification. Below we present the examples with bilingual support.

Abbreviation expansion. Figure 3 contrasts correct vs. incorrect abbreviation handling.

Medication explanations. Figure 4 shows how therapeutic purpose explanations improve treatment comprehension.

6. Discussion

This section evaluates the system’s performance, analyzes error patterns, and identifies future development directions.

6.1. System Performance

SimpliMED demonstrates the effectiveness of a modular pipeline and domain-adapted Small Language Models (SLMs). Our fine-tuned models for abbreviation detection and medication recognition achieved F1 scores of 0.90 and 0.91, respectively, proving that specialized models can be both efficient and accurate. Integrating validated resources like SEDOM and AEMPS reduces hallucinations,

Abbreviation Expansion Comparison

Note: Abbreviation handling is optimized for Spanish.

Spanish:

Original: “Antecedentes de HTA y FA permanente. Ecocardiograma con FEVI preservada.”

Without module: “Antecedentes de **Hiper-Tensión Área** y **Factor de Angustia** permanente. Ecocardiograma con **Factor Evidente** preservada.”

With module: “Antecedentes de **Hipertensión Arterial (HTA)** y **Fibrilación Auricular (FA)** permanente. Ecocardiograma con **Fracción de Eyección del Ventrículo Izquierdo (FEVI)** preservada.”

English:

Original: “History of HT and permanent AF. Echocardiogram showing preserved LVEF.”

Without module: “History of **Hyper-Tension Area** and permanent **Anguish Factor**. Echocardiogram with **Evident Factor** preserved.”

With module: “History of **Arterial Hypertension (HT)** and permanent **Atrial Fibrillation (AF)**. Echocardiogram showing preserved **Left Ventricular Ejection Fraction (LVEF)**.”

Figure 3: Without module: incorrect expansions. With module: correct clinical meanings.

while expert evaluation by cardiologists provides deeper insights than standard readability metrics. However, a key limitation is the lack of direct patient testing, which remains necessary to fully validate comprehension improvements.

The use of *LLaMA 3.1 8B* involves a trade-off between quality and computational cost. While requiring significant GPU resources, this model size offers a balance that smaller models cannot yet match. Furthermore, our section-based segmentation approach ensures consistency and prevents information loss, highlighting the importance of structural awareness in clinical text processing.

6.2. Error Analysis

Critical error patterns included incorrect abbreviation disambiguation (e.g., confusing portal and pulmonary hypertension) and incomplete explanations for combination medications. Occasional factual hallucinations in complex clinical interpretations (like Tako-Tsubo syndrome) underscore the limitations of current model reasoning in technical sections.

Future iterations will prioritize Retrieval-Augmented Generation (RAG) to ground outputs in

Treatment Section with Therapeutic Explanations
Spanish: Original: "Apixaban 5mg 1-0-1; Aspirina 100mg 1-0-0; Atorvastatina 80mg 0-0-1" With treatment module: • Anticoagulación: Apixaban 5mg. Tomar por la mañana y por la noche. Sirve para prevenir coágulos. • Antiagregante: Aspirina 100mg. Tomar por la mañana. Sirve para proteger las arterias. • Lípidos: Atorvastatina 80mg. Tomar por la noche. Sirve para bajar el colesterol y proteger el corazón.
English: Original: "Apixaban 5mg 1-0-1; Aspirin 100mg 1-0-0; Atorvastatin 80mg 0-0-1" With treatment module: • Anticoagulation: Apixaban 5mg morning and evening—prevents clots. • Antiplatelet: Aspirin 100mg morning—protects arteries. • Lipids: Atorvastatin 80mg evening—lowers cholesterol and protects the heart.

Figure 4: Example of medication simplification. The treatment module categorizes medications and adds therapeutic purpose explanations.

verified knowledge bases and clinical knowledge graphs. Additionally, implementing a human-in-the-loop validation step will ensure patient safety during the transition toward fully automated systems.

6.3. Evaluation Considerations

Our pragmatic evaluation combines automatic metrics with expert review, addressing the limitations of readability indices which often ignore semantic fidelity. Despite the small sample size, expert feedback correctly prioritizes clinical accuracy and patient safety.

We acknowledge the potential for bias when using LLM-generated annotations as reference data. While manual review mitigates major errors, this "circular" evaluation highlights the need for expert-annotated benchmarks specifically for the Spanish clinical domain.

6.4. Future Work

Future research will focus on extending the system to other specialties, developing automated metrics tailored for medical simplification, and conducting

direct patient comprehension studies through interactive interfaces.

7. Conclusions

We presented SimpliMED, a modular system for simplifying cardiology discharge reports using Large Language Models and specialized NLP components. Our approach addresses access barriers that prevent patients from understanding their medical documentation through an architecture that integrates three complementary components.

The section-based processing approach preserves clinical completeness by applying tailored prompts to each report section, avoiding the information loss observed with global processing approaches. The abbreviation handling component, implemented through fine-tuned detection achieving F1 of 0.90 combined with contextual disambiguation, significantly improves upon LLM-only approaches that frequently misexpand medical abbreviations with potential clinical consequences. The medication explanation module transforms cryptic prescription lists into categorized, purposeful treatment summaries through integration of the AEMPS therapeutic dictionary containing 14,611 medications.

Expert evaluation by two cardiologists yields 75% overall quality, with strong performance on admission reason sections at 91% and current illness at 75%, while identifying challenges with technical sections like diagnostic tests at 44% that require clinical interpretation beyond current model capabilities.

While demonstrating significant potential, SimpliMED requires further development before clinical deployment. Critical needs include enhanced disambiguation mechanisms to prevent dangerous misinterpretations, clinical validation pipelines to catch conceptual errors, and robust handling of complex medical concepts. The expert feedback particularly emphasizes restructuring outputs toward direct, patient-centered communication rather than technical reformulation.

This work represents a meaningful step toward improving patient access to medical information and reducing health literacy barriers in Spanish healthcare contexts. As legal requirements for accessible institutional communications extend to the medical domain, systems like SimpliMED will become increasingly essential for ensuring patients' right to understand their own health information.

8. Limitations

Several limitations constrain the interpretation and generalization of our results. First, expert evaluation covered only 8 reports due to resource con-

straints; while informative, this sample size limits generalization of quality assessments and necessitates larger-scale clinical validation before deployment. Second, no established automatic metrics exist for medical text simplification quality, traditional readability indices do not account for semantic fidelity or clinical accuracy, complicating iterative development and benchmark comparisons.

The system was developed and evaluated exclusively on cardiology reports. Generalization to other medical specialties requires additional adaptation and validation, as terminology, abbreviations, and section structures vary significantly across clinical domains. Furthermore, while expert evaluation assesses clinical accuracy, we did not conduct comprehension testing with actual patients, leaving the ultimate measure of simplification success, improved patient understanding, to be validated in future work.

9. Ethical Considerations

This research was conducted in accordance with the ethical principles of the Declaration of Helsinki and was approved by the corresponding institutional Ethics Committee (code: SICEIA-2024-001925). Informed consent was waived by this committee due to the retrospective design and the use of anonymized data from hospital discharge reports.

The study complied with applicable data protection regulations, including the European General Data Protection Regulation and Spanish healthcare data protection laws. The corpus of 307 cardiology discharge reports was fully anonymized prior to use, with all patient identifiers, dates, and potentially identifying information removed by the providing medical institution.

The system is designed as a clinical decision support tool and is not intended to replace professional medical judgment. Simplified reports are meant to complement, not substitute, direct patient-physician communication. We acknowledge the risk of oversimplification potentially omitting clinically relevant nuances, and recommend that any deployment include physician review of simplified outputs before patient delivery.

10. Acknowledgements

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded

by the European Union under the Citizens, Equality, Rights and Values programme, Project CONSENSO (PID2021-122263OB-C21) Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP_PIDI_2024_00852:) funded by Junta de Andalucía

11. Bibliographical References

- Mark Adler. 2012. [The plain language movement](#). In Lawrence M. Solan and Peter M. Tiersma, editors, *The Oxford Handbook of Language and Law*. Oxford University Press.
- AENOR. 2018. *Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos*. AENOR INTERNACIONAL SAU, Madrid.
- Lydie Bednarczyk, Daniel Reichenpfader, Christophe Gaudet-Blavignac, Amon Kenna Ete, Jamil Zaghir, Yuanyuan Zheng, Adel Bensahla, Mina Bjelogrić, and Christian Lovis. 2025. [Scientific evidence for clinical text summarization using large language models: Scoping review](#). *J Med Internet Res*, 27:e68998.
- Robert-Mihai Botarleanu, Mihai Dascalu, Scott Andrew Crossley, and Danielle S. McNamara. 2020. [Sequence-to-sequence models for automated text simplification](#). In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II*, pages 31–36, Berlin, Heidelberg. Springer-Verlag.
- Stefan Bott and Horacio Saggon. 2012. [Automatic simplification of spanish text for e-accessibility](#). In *Computers Helping People with Special Needs*, pages 527–534, Berlin, Heidelberg. Springer Berlin Heidelberg.
- 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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

- Leonardo Campillos-Llanos, Ana Rosa Terroba Reinares, Sofía Zunder Puig, Ana Valverde-Mateos, and Ana Capllonch-Carrión. 2022. [Building a comparable corpus and a benchmark for Spanish medical text simplification](#). *Procesamiento del Lenguaje Natural*, 69:189–196.
- John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*.
- Joshua C. Denny, Anderson Spickard, Kevin B. Johnson, Neeraja B. Peterson, Josh F. Peterson, and Randolph A. Miller. 2009. [Evaluation of a method to identify and categorize section headers in clinical documents](#). *Journal of the American Medical Informatics Association*, 16(6):806–815.
- Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2023. [Multilingual simplification of medical texts](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16662–16692, Singapore. Association for Computational Linguistics.
- Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. *AMIA ... Annual Symposium proceedings / AMIA Symposium*. AMIA Symposium, 2010:366–70.
- J. Richard Landis and Gary G. Koch. 1977. [The measurement of observer agreement for categorical data](#). *Biometrics*, 33(1):159–174.
- OECD. 2023. [Survey of adult skills \(PIAAC\): 2023 country note – Spain](#).
- Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. [A survey of automated methods for biomedical text simplification](#). *Journal of the American Medical Informatics Association*, 29(11):1976–1988.
- 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.
- Paul Rust, Julian Frings, Sven Meister, and Leonard Fehring. 2025. [Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations](#). *Communications Medicine*, 5.
- Laura Vázquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2021. [Investigating text simplification evaluation](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 876–882, Online. Association for Computational Linguistics.
- Laura Vázquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2023. [Document-level text simplification with coherence evaluation](#). In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 85–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- 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. Curran Associates, Inc.
- Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. [Sentence simplification by monolingual machine translation](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.
- William Xia, Ishita Unde, Brian David Ondov, and Dina Demner-Fushman. 2025. [JEBS: A fine-grained biomedical lexical simplification task](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17654–17666, Vienna, Austria. Association for Computational Linguistics.
- 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.

12. Language Resource References

- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. [Mlsum: The multilingual summarization corpus](#).

Dutch Metaphor Extraction from Cancer Patients' Interviews and Forum Data Using LLMs and Human in the Loop

Lifeng Han^{1,2}, David Lindevelt², Sander Puts³
Erik van Mulligen⁴, Suzan Verberne²

On behalf of 4D PICTURE Consortium

¹Biomedical Data Sciences, Leiden University Medical Center, NL

²Leiden Institute of Advanced Computer Science (LIACS), Leiden University, NL

³Department of Radiation Oncology (MAASTRO),

GROW - Research Institute for Oncology and Reproduction,

Maastricht University Medical Centre+, Maastricht, NL

⁴Department of Medical Informatics, Erasmus University Medical Center Rotterdam, NL

{l.han, s.verberne}@liacs.leidenuniv.nl

Abstract

Metaphors play an important role in communication between cancer patients and clinicians. In this work, we present the first study on automated extraction of Dutch metaphors from cancer patient interviews and online forum data using large language models (LLMs) and human validation. We develop a human-in-the-loop extraction framework combining structured prompting, automatic verification, and expert linguistic evaluation. We evaluate multiple open-source LLMs and prompting strategies, including chain-of-thought and knowledge-guided prompting. Our approach resulted in the HealthQuote.NL dataset, containing 130 validated metaphors. We analyse common model errors, including hallucination and confusion with idioms, and show that refined prompting achieved 63.2% precision. This work provides both a novel dataset and an extraction framework to support patient-centred healthcare communication and future metaphor research. We share our prompts and resources at <https://github.com/4dpicture/HealthQuote.NL>

Keywords: Metaphor, Healthcare, Cancer Patients, Dutch language, LLMs

1. Introduction

According to the Contemporary theory of metaphor, a metaphor is “a cross-domain mapping in the conceptual system” (Lakoff, 1993). Metaphor can be expressed in different linguistic units, such as words, phrases, or sentence levels. People use metaphors in daily life with or without realising them.

In the healthcare domain, metaphors can play a helpful role in the communication between caregivers and cancer patients (Semino et al., 2017b; Harrington, 2012; Liu et al., 2024): patients can be helped by comparing their cancer treatment process to e.g. journeys or battles. In the UK, the Metaphor Menu has been a successful project, publishing a pamphlet of potentially helpful metaphors for patients living with cancer.¹

In this paper, we take the first steps in collecting a set of Dutch-language metaphors related to living with cancer. We extract these metaphors from patient utterances, both spoken (transcribed) and written. To this end, we analyse two distinct corpora of Dutch patient narratives: interviews and blog posts (kanker.nl). We experiment with a suite of small, local LLMs and a range of prompting techniques to extract metaphors from texts. After analysing the

output with simple instruction prompts, we identify 7 challenges for this task, including hallucination of metaphors, figurative language or idioms mistaken as metaphors, and abstraction instead of extraction of the originally mentioned metaphors. With more advanced prompting techniques, including chain-of-thought and iterative self-prompting, we were able to address these challenges to some extent and extract useful metaphors from the two datasets.

In summary, our contributions are:

- We present the first study investigating large language models for metaphor extraction in Dutch cancer patient narratives to support patient-centred healthcare communication.²
- We develop a human-in-the-loop extraction framework combining structured prompting, automatic verification, and expert linguistic validation.
- We construct HealthQuote.NL, a curated dataset of 130 validated Dutch metaphors derived from interviews and forum data.

¹<https://wp.lancs.ac.uk/melc/the-metaphor-menu/>

²Regarding data sharing of HealthQuote.NL, to protect patients' privacy, we generate paraphrased and synthetic examples using the metaphors we extracted from online forum text.

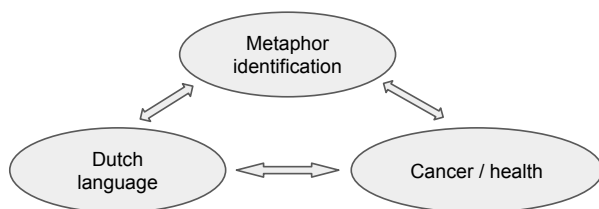


Figure 1: Related Work in Three Dimensions

- We provide an empirical analysis of prompting strategies, identifying common failure modes such as hallucination, abstraction, and idiom confusion.

2. Related Work

In this section, since we did not find literature work exactly on metaphor identification for Dutch cancer patients, i.e. satisfying the triple (metaphor, Dutch, cancer), we introduce works from three perspectives: metaphor identification in Dutch, metaphors in cancer, and metaphors in Dutch healthcare, as in Figure 1.

2.1. Metaphor Identification in Dutch

Regarding metaphor identification and extraction methods in Dutch, to the best of our knowledge, little work has been done. One PhD thesis by Pasma (2011) focused on metaphor and register variation in Dutch news discourse. They extended the metaphor identification procedure (MIP) from Group (2007) by adapting it to some new situation when “cross-domain mapping” appears for certain words, including “direct metaphor”, “implicit metaphor”, and “metaphor flag”. This adaptation has been applied to the Dutch discourse corpus on news texts. The MIP principle is a symbolic text examination on a word-by-word basis. Later on, there have been a few works published as book chapters (non-open-access books) by the same research group, following the PhD thesis project, including (Pasma, 2012, 2019).

Another loosely related work is from De Knop and Perez (2014), which studies the conceptual metaphors of Dutch and German for efficient teaching and, alternatively, aims at verifying them.

Our work will be the first exploring and investigating LLMs for Dutch metaphor identification and extraction.

2.2. Metaphors in Cancer

Various researchers have conducted metaphor studies within the cancer domain. Review articles in this topic include Harrington (2012); Liu

et al. (2024). Semino et al. (2017a) carried out comparative studies on metaphors used by cancer patients vs health professionals at online forums in the UK between 2007 and 2013. It included 56 patients and 307 health professionals. The study showed that “violence” and “journey” related metaphors are not necessarily negative or positive by default. It depends on how they are used. The “violence” metaphors were further studied and discussed in their following work on communicating nuanced results in Language Consultancy (Demjén and Semino, 2020). Similar work was conducted by Declercq et al. (2023) on metaphors use of machines, journeys, and prisons for chronic pain patients in consultations to describe pain, illness, and medicines.

There are courses designed, e.g. “Cancer as Metaphor” by Penson et al. (2004) for explaining the strengths and weaknesses of using metaphor in patient-caregiver relationships. The instructions covered different roles, including facilitator, oncologist, psychiatrist, and psychologist.

Focusing on the staging, researchers studied “imprisonment”, “burden”, “battle”, and “journey” metaphors on patients living with *advanced cancer* using blog posts (Hommerberg et al., 2020), and a focus group design with eighteen people on cancer survivor (Appleton and Flynn, 2014).

From a gender perspective, Gibbs Jr and Franks (2002) studied the relation between language use and thoughts from 6 females who live with cancer. From cancer types, Pfeifer et al. (2024) conducted a study on how metaphors can help the communication of information exchange from breast cancer patients about depressive symptoms.

There are also focused studies on language or region-specific, e.g. Spanish (online forum) (Magaña and Matlock, 2018), the UK (corpus-based) (Semino et al., 2017b), American and Nigerian (self-help books) on “military, journey, personification, and sports” (Nwankwo, 2024).

Most work has focused on one source of data, e.g. either online data, interviews, or books. However, we aim to *explore the difference between verbal and online-written data*. Similar to (Semino et al., 2017b), in the interview data we have, there are 3 roles (patient, significant other, and interviewer), but in Dutch (the Netherlands) instead of English (UK).

2.3. Metaphors in Dutch Healthcare

To our best knowledge, there is no prior work on Dutch-language metaphors in the cancer domain. However, there is some work on metaphor-related research in other Dutch healthcare domains that we list here.

To better understand the stigma toward *dementia* patients, Creten et al. (2022) conducted sentiment

analysis of Dutch language using tweets from 2019 to 2020 into seven dimensions, including joke and metaphor, in addition to others such as information, organization, personal experience, politics, and ridicule. The authors identified 10% of the tweets were metaphorical during manual coding.

To study the needs of *hypertensive* patients expressing themselves using metaphors, analogies, and symbols, such as the causes and consequences of the illness, Schuster et al. (2011) used 55 patients of three ethnic groups in the Netherlands, including Creole, Dutch, and Hindustani descent, for in-depth interviews. The finding shows that patients use machine and enemy-related metaphors to explain the origins and consequences of hypertension, with its characteristics of silence, invisibility, and uncontrollability.

In another study by Hillen et al. (2025), the authors investigated the usage of metaphoric language by clinicians for family members of *ICU* patients. They sampled 36 neonatal, pediatric, and adult patients, together with 104 of their family members. Using the transcribed and anonymised audio recording of conversations, the findings show that metaphors were used frequently and predominantly by clinicians to share clinical information and the awareness of using metaphors by clinicians can improve effective information exchange. However, the authors also discussed that both harms and benefits arise from clinicians using metaphors, e.g. shared decision making (SDM) might be hindered if clinicians actively and consciously use metaphor to steer the families into a direction they consider as the best for patients.

In contrast to this, in our work, we aim to *identify and extract metaphors used by patients and their family members* (significant others), for better support of patient care.

3. Methods Design and Development

To identify metaphors used in texts in patient records, we explore different kinds of techniques on prompting LLMs, including Persona (Lutz et al., 2025), chain of thought (CoT) (Saparov and He, 2023), iterative self-prompting (ISP, aka LLM-assisted prompting) (Romero et al., 2025; Ren et al., 2025), and knowledge-inserted prompting (Dong et al., 2024). To analyse if the structure, size of LLMs, and their pre-trained domain knowledge matter, we use LLMs from different sizes and structures, as well as domain-specific ones, including models pre-trained on medical knowledge.

As shown in Figure 2, the initial prompt we used is the "Instruction Prompt (LInP)" using Persona, role of "system" and "user", without CoTs. Messages we used for prompting are:

- "role": "system", "content": "You are an assistant that processes and analyses documents."
- "role": "user", "content": "Here is the document: <document-content> Please tell me metaphors that people used in the conversation in this document and list them in bullet points. Can you offer the English translation version too? Can you also offer the full sentences in which the metaphors were used?"

Then, in the refined prompts (RPs), we added CoT and ISP by asking LLMs to generate prompts for such a task. After a few iterations of refinement, the LLM-assisted long and structured prompt message for this task includes the following key components.

- Persona: an expert linguist
- Strict Extraction Protocol
 - Document Scanning
 - Metaphor Identification, CoTs with examples
 - Verification Requirements, different levels
 - Data Extraction Format
 - Structured Output (Population)
- Quality Control Rules
- Final Instructions

Among the Refined Prompts (RP v1 and v2), the difference is that for v1, we only gave three simple and short metaphorical sentences for few-shot prompting:

- "This treatment is a journey" (medical process → travel)
- "Fighting cancer" (disease → war)
- "The tumor is growing like a weed" (cancer → plant)

For v2, we inserted the entire English Metaphor Menu list (17 from the public website [Metaphor Menu, Lancaster University \(2025\)](#)) with our categorisation. The different settings for RP-v1 and RP-v2 are to investigate 1) if the English Metaphor Menu can guide the LLMs to extract better metaphors; 2) if the English Metaphor Menu will introduce bias to LLMs to extract only metaphors like the English metaphors.

For automatic verification (auto.verify), we implemented an external checklist tool for LLMs: if they are able to find the exact text in the document that expresses the metaphor, identify the exact section, identify the speaker's role (when there are multiple roles), and identify the metaphor (not literal medical terms).

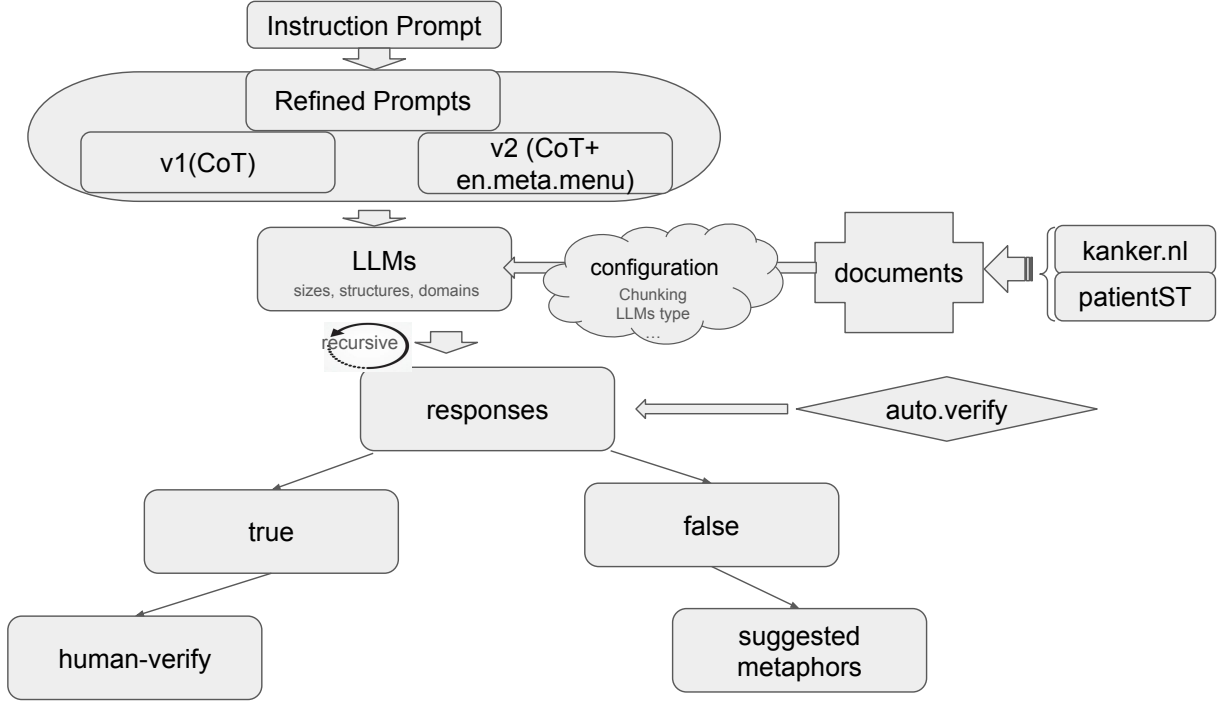


Figure 2: HEALTHQUOTE.NL: Extraction Framework using LLMs and Human in the Loop.

For extracted metaphors, we ask LLMs to group them into 1) types: word, phrase, sentence, and extended, 2) source domain: violence, journey, nature (garden), games, music, fairground, unwanted guests, religion, control, building, machine, and other, and 3) function in context: explanation, coping, empowerment, relationship, prognosis, treatment, emotion, and humour.

For security and privacy reasons on processing patients’ data, even though they are already anonymised, we prioritise to use local LLMs through the ollama platform³ over commercial models. The open-source models we used include both medical and general domains.

4. Experiments

4.1. Data

For **Patient Story Telling (patientST) Interview data**, we used the transcriptions of 13 documents from 13 interviews in an oncology setting obtained from (Griffioen et al., 2021). There are three roles for each interview: cancer patient, a family member or friend of the patient (significant other), and interviewer (researcher). They are specified as P (Patiënt), N (Naaste), and O (Onderzoeker) in the free text. The largest document has 13,777 words and 32 pages, while the smallest document has 5,596 words and 16 pages. See Figure 3 for statistics per document.

³<https://ollama.com/library>

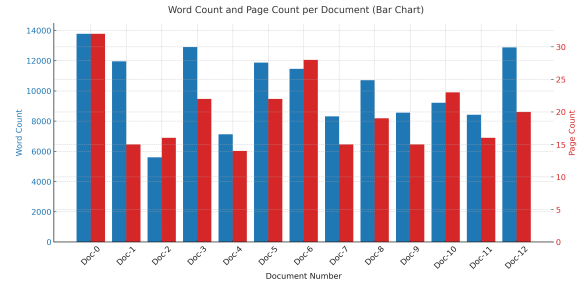


Figure 3: Interview Data Statistics.

For **cancer patient forum data**, we have a collection in Dutch of 15,653 blogs, 17,290 comments, 2,246 group discussions, 5,777 “ask a professional” items in which experts answer questions from patients, and 10,134 reactions on “ask a professional” from three cancer types: breast cancer (3,524), prostate cancer (2,613), and melanoma (614). The patient content includes treatment phases, treatments, and long-term consequences.

We have approval to use both the two sets of data for research. We can not disclose the original forum text. Considering this, instead of displaying the original text, we will give paraphrases or synthetic example texts generated from the extracted metaphors.

4.2. Model Parameter Settings

The models, parameters, and their values we explored (model configuration) are:

- **Prompts:** Instruction Prompt (IP), RP-v1, RP-v2
- **Text split:** max-token=4000, overlap-token: 40, char-per-token: 4
- **Models:** qwen3:8b, gemma3:12b, gemma3:27b, llama3.1:8b, mistral:7b, deepseek-r1:8b, meditron:7b, medllama2:7b
- **Options:** context size window=32768, temperature=0.8, top-k=40, top-p=0.9

Among these parameters, higher temperature allows LLMs to generate more diverse or creative outputs, but also increase the chance of nonsensical, irrelevant, or factually incorrect outputs. We empirically verified that this did not significantly reduce extraction faithfulness due to human validation. Although metaphor extraction is primarily an extractive task, a moderately high temperature was used to allow models to detect less explicit metaphorical expressions that may not follow fixed linguistic patterns. Human validation ensured that only faithful extractions were retained.

4.3. Evaluation Setting

In addition to the automatic verification steps shown in Figure 2, we conducted a human qualitative evaluation of the extracted metaphors. Three native Dutch speakers with PhD degrees and expertise in computational linguistics, health communication, or related fields independently reviewed the outputs. Each candidate metaphor was assessed for **faithfulness** (explicitly present in the original text, not hallucinated or paraphrased), **metaphoricity** (genuine cross-domain conceptual mapping rather than literal language or idioms), and **contextual appropriateness** (accurately reflecting the intended meaning in context). Each item was labelled as *valid* or *invalid metaphor*, and disagreements were resolved through discussion to reach consensus. This human-in-the-loop evaluation ensured linguistic validity and contextual correctness, and enabled the construction of the HealthQuote.NL dataset for downstream analysis and clinical communication applications. Only validated metaphors were included in the final corpus and reported in the results.

4.4. Quantitative Evaluation

To assess the effectiveness of the prompting strategies, we measured the precision of metaphor extraction under different prompt settings. Precision

was defined as:

$$Precision = \frac{Metaphor_{validated}}{Metaphor_{total}}$$

where $Metaphor_{validated}$ and $Metaphor_{total}$ are the number of metaphors validated after human verification and the total number of candidate metaphors generated, respectively.

Table 1 shows the comparison between Instruction Prompt and Refined Prompts.

Prompt	Generated	V.	P.
Instruction Prompt	72	41	56.9%
Refined Prompt v1	38	24	63.2%
Refined Prompt v2	174	24	13.8%

Table 1: Comparison of metaphor extraction performance across prompting strategies (V: valid; P: precision)

Using the initial instruction prompt, the model generated 72 candidate metaphors across 13 interview documents. After human verification, 41 were confirmed as valid metaphors, resulting in a precision of 56.9%. The remaining candidates included hallucinated metaphors, paraphrased expressions, and idioms.

The refined prompt (RP-V1) achieved substantially higher precision, confirming that structured prompting, chain-of-thought reasoning, and explicit extraction constraints reduce hallucination and improve faithfulness.

The refined prompt v2 generated substantially more candidate metaphors (174) due to broader model testing, but only 24 were validated as true metaphors, resulting in a precision of 13.8%. This indicates that while providing the full English Metaphor Menu increased sensitivity, it also introduced significant *noise* and *over-interpretation* for some LLMs. In contrast, refined prompt v1 achieved the best balance between precision and extraction quality.

4.5. Output with Initial Instruction Prompt

We first list some challenges with example outputs using llama3.1:8b and only initial Instruction Prompt (I.InP) without CoT on interview data:

1) Some **abstractive or paraphrased** summarised metaphors, instead of extractive: e.g. a) from Doc-6: "Een strijd voeren voor je lijf" (Fighting a battle for your body) from Full sentence: "Dat doe je ook deels voor je familie. Want die vinden, ja die willen graag dat je blijft en dan zegt het lijf vanzelf 'ja ik ga knokken'." (You do it also partly for your family. Because they want to stay and then your body says itself "yes I'll fight"); and another one b) from Doc-12: "Het is alsof je in de nacht ziet en je

weet niet wat er aan de hand is" (It's like seeing things in the dark and not knowing what's going on) from Full sentence: "Maar het was alsof ik in de nacht stond en niets zag. Ik wist niet wat er aan de hand was." ("But it was like standing in the night and seeing nothing. I didn't know what was going on.")

2) **Fail to translate** simple sentences: e.g. from Doc-8: De zonsondergang van zijn leven" ("no direct English translation" from LLM, shall be "the sunset of his life") but it gave an explanation of the metaphor: "using sunset as a metaphor for the end of one's life".

3) **Suggesting metaphors output** plain text: e.g. from Doc-9: For the original text in Dutch: "Het is heel belangrijk dat patiënten heel positief blijven en de familie ook - niet alleen familie, vrienden-meepraten en vooral niet bang zijn" (translation in English: "It's very important for patients to stay positive and for the family too - not just family, friends - to discuss things and especially not be afraid"), LLM suggests the following metaphor: "Comparing the patient's journey to a ship navigating through stormy waters. The speaker emphasises the importance of staying positive and discussing things openly.", which can be useful in communication to support patients illustratively.

4) **Figurative language** rather than strict metaphors: e.g. in Doc-9: the original Dutch text "Ik denk dat we allemaal een beetje te vroeg geboren zijn" (translation in English: "I think we're all a bit too early born"). It was interpreted as metaphor by LLM "Comparing human beings to products that are not yet ready for use. The speaker humorously suggests that they and their family members are still learning how to navigate the healthcare system." However, this is more as a figurative language expression, instead of metaphor mapping A to B.

5) **Over-Interpreting metaphors**: e.g. from Doc-10, a metaphor "een klein schip in de storm" (English Translation: "A small ship in a storm") is suggested from "het wachten" ("the waiting"), although the context did not say much about this metaphor. Instead, the researcher was asking the patient about the concern of process, as quoted below. "O: Heel goed, dank u. Wat in het hele traject vond u eng of vervelend? P: Het wachten. O: Wat waren met name de momenten van het wachten?" (translation: "O: Very good, thank you. What did you find scary or annoying about the whole process? P: The waiting. O: What were the specific moments of waiting?")

Despite these issues, we were able to extract and collect 28 metaphors from the 13 interviews after human verification using the initial instruction prompt. They reflect some of the source domains we defined later for the refined prompts, including journey, religion, control, machine, nature, and vio-

lence. From a functional perspective, they can help communicate: feeling, hospital, process, dealing, life. These also reflect our function design in RPs (Section 3).

We also explored I.inP with verification tool (I.inP.VT) control and the original English Metaphor Menu webpage link (I.inP.V.L). Overall, 7 were re-identified by the model, however, there are also 3 variations and 13 new ones generated, together with the original 28 resulting in **41 in total from the initial instruction prompts**. For examples of variations, e.g. from Doc-1, the full sentence "P: Ja, die morfine maakt dat je, die die heeft me ook wel angstig gemaakt." ("P: Yes, that morphine makes you, that made me anxious too."), I.inP generated "Morfine brengt angstgevoelens naar boven" (Morphine brings anxiety feelings to the surface), while the I.inP.VT generated more indirect and figurative expression "Er zit een morfine achter je" (There's morphine behind you). The metaphor is that the morphine is a physical presence that's affecting their behaviour.

In the outputs using the verification tool and the English Metaphor Menu website page link, we have new findings that we list here following the aforementioned list.

6) The LLM extracted **idioms instead of metaphors**. For instance, from Doc-1 "Een baby-hoeveelheid" (A baby-amount) is extracted to express the dietary change from the patients, from the full sentence: "Het is maar een baby-hoeveelheid. Nog niet eens de helft van hij vroeger at." ("It's only a baby's amount. Not even half of what he used to eat.")

7) The LLM **hallucinated metaphors**, e.g. from Doc-5, LLM generates a metaphor "De dokter is een dominee" (The doctor is a preacher), from the original sentence: "Die arts-assistent was echt 0 komma 0." (English translation: "That doctor's assistant was 0 point 0."). LLM explains "The speaker compares the doctor to a preacher, implying they were ineffective and unhelpful." Firstly, it switched the role from "doctor's assistant" to "the doctor". Secondly, "preacher" was not used in the context. The difference between Point 5 and Point 7 is that, Over-interpreting metaphors is more than suggestive metaphors but still try to interpret the situation; however, Hallucinated metaphors even misinterpreted the roles in the context, e.g. "doctor" to the "assistant".

4.6. Output with Refined Prompts

We applied Refined Prompts (RPs) with CoTs (v1 and v2) and English metaphor menu as inserted knowledge (full text, instead of webpage link), in-context learning for v2. We applied these two RPs on the interview data, and the collected metaphors

are categorised into types (word, phrase, sentence), source domains, and functions.

With RPs and inserted knowledge, **24 metaphors** are collected after human verification from collective LLMs, Table 2. The output metaphors from RPs are more closely related to the original context and have less hallucination. Thus, in total, the initial instruction prompts and RPs have **collected 65** (41+24) metaphors from interview data.

However, it is still not possible to totally avoid it. For instance, from Doc-3, LLMs (mistral:7b) attempt to create a metaphor out of the original phrase "Volle melk, volle yoghurt.." ("Full milk, full yoghurt..."). It describes "The patient compares their current state to a full milk or yogurt container, possibly emphasising the feeling of being overwhelmed.", with "reasoning": "This is metaphorical because the context suggests an emotional burden that goes beyond physical nourishment." However, the original context has only talked about food.

These findings are also supported by recent work from (Fuoli et al., 2026) where the authors discussed the hallucinations and prompt sensitivity of LLMs on metaphor detection. This justifies the human in the loop design of our work. In addition, they also carried out fine-tuning LLMs on this task, mentioning grey areas and conceptual challenges in metaphor theory.

4.7. Application to Blog Data

As a pilot application, we applied our automatic metaphor extraction method (Gemma3:27b via Olama) to the first 100 blog posts on Kanker.nl. The method is designed to retrieve health-related metaphorical expressions by construction, and thus targets figurative language specifically linked to illness experience rather than general metaphor use. The output included both novel metaphors and frequent idiomatic expressions that are part of everyday Dutch. Examples of conventionalised or dead metaphors include "te horen krijgen" ("to be told"), "meten is weten" ("to measure is to know"), "ik baal als een stekker" ("I'm as annoyed as a plug"), and "laten we hopen dat alles met een sisser afloopt" ("let's hope it all blows over quietly"). While emotionally expressive, such idioms are generic and not specific to the cancer experience.

Across the 100 posts, the model identified 39 posts containing at least one candidate expression, yielding 65 distinct health-related metaphor instances. All extracted instances were manually reviewed with respect to their health-related relevance and potential downstream utility. Of the 65 instances, 10 (15.4%) satisfied the stricter criteria of novelty, contextual specificity, and therapeutic relevance required for potential inclusion in a metaphor menu. This proportion reflects the downstream

acceptance rate under the defined application criteria and should not be interpreted as a precision estimate for metaphor detection. Expressions not selected at this stage were excluded due to conventionality or limited conceptual richness rather than extraction error.

These most salient examples demonstrate how patients and relatives employ creative language to articulate emotion, reframe vulnerability, and construct meaning in relation to illness. The selected metaphors, grouped below by their metaphorical vehicle, illustrate the method's ability to surface conceptually rich, health-related figurative expressions in naturally occurring patient discourse.

The Party Here, cancer is personified as an uninvited celebrant. The party metaphor conveys irony and defiance, transforming fear into dark humor. This creative reversal allows the writer to express anger and regain narrative control by reframing the illness as something absurd rather than purely tragic.

"I was told not to use worrying titles anymore, so this time I've decided to make it a celebration instead – my cancer is having a party of its own."

The Car The damaged car serves as a vivid metaphor for physical and emotional wear. It conveys the experience of being battered but not beyond repair, blending vulnerability with resilience. The detailed imagery grounds the illness experience in tangible, everyday language that conveys fatigue and determination simultaneously.

"I feel like an old wreck of a car, the kind that can't be patched up with filler anymore – it needs entirely new bodywork, like the rusty 2CV sitting in someone's garden."

The Lighthouse The lighthouse represents guidance, stability, and safety. It evokes the dependable presence of loved ones who provide light and orientation amid uncertainty. This metaphor articulates the relational and emotional support structures crucial for coping with illness.

"A lighthouse – someone solid and constant, always there to guide you through the dark."

The Storm The storm metaphor externalizes psychological turmoil and intrusive thoughts. By likening mental distress to turbulent weather, the speaker gives form to anxiety and confusion, making inner chaos visible and easier to communicate.

"It feels as if a wild storm is roaring through my head."

Overall	Qwen	Mistral	MedLlama	DeepSeek	Llama3.1:8B	Gemma3:12B	Gemma3:27B
24	5	8	1	0	5	1	4

Table 2: Validated metaphor counts from collective LLMs using RPs on interview data.

The Train The train functions as a metaphor for relentless motion and loss of control. These images capture the unstoppable pace of medical treatment and the sense that life continues regardless of one’s readiness, balancing fatigue with perseverance.

“Just when you think you’ve stepped off one speeding train, you realize you’re already on the next one.”

“Life keeps pounding along like a fast-moving train – whether I’m ready to get on or not.”

The Painting Referring to the red surgical markings before an operation, the painting metaphor reinterprets the medicalized body as a canvas. It reframes vulnerability as something that can hold meaning and even beauty, allowing patients to reclaim agency over their altered bodies.

“My stomach looks like a work of art now, covered with careful red lines and circles drawn before surgery.”

The Final Chord The musical metaphor of the *final chord* transforms death into a form of completion rather than cessation. It provides emotional and narrative closure, allowing grief to be expressed through harmony and resolution.

“This feels like the final note in the song of our journey together – the moment of closure that sends everything gently into silence.”

These selected metaphors show how writers on *Kanker.nl* employ creative figurative language to convey themes of control, damage, support, turbulence, transformation, and closure. The combination of automatic extraction and qualitative interpretation reveals how metaphor functions as a key tool for emotional expression and meaning-making in online narratives of illness.

4.8. GPT on Forum Data

To compare with the open-source LLMs, we also used GPT models from OpenAI on the forum data for the metaphor extraction task and prompted GPT5 to map the extracted metaphors to the English Metaphor Menu and provide a confidence score. We list some of our extracted metaphors using GPT5 (*paraphrased*) in the appendix section, which align with the original English metaphor menu, as in Figures 4 and 5.

5. Discussion

Firstly, without the CoTs and verification tools we implemented, the initial instruction prompts produced abstractive and suggestive metaphors, paraphrased from what was said in the interview data, instead of word-for-word extractions. With the refined prompts and the inserted English metaphor menu knowledge, the LLM-generated metaphors are more closely reflecting the transcription content, but also less diverse or creative than the initial prompting outputs.

Regarding the corpus **HealthQuote.NL** we generated, it can be useful for multiple purposes, as a whole or a subset of columns. For instance, users can focus on the Dutch metaphors column for Dutch-focused patient care support. The Dutch-English bilingual metaphors can be used for non-Dutch native speakers but who speak English in the care setting. The categorisation, description, and reasoning of metaphors can be used to conduct a metaphor study itself and the relations within. The prompts we used and the corresponding varied outputs from LLMs can be explored by AI and NLP practitioners who are interested in exploring prompt engineering for better analysis of non-literal language use.

For the phrase-level and word-level metaphors, we always add the context sentence as examples to the dataset for understanding and usage. For instance, at word level, the pair of original quote “balletje” and translation “ball”, is accompanied with the context “O: dan gaat het balletje heel snel rollen.” (*O: Then the ball starts rolling very quickly.*) We also give the description “In this case “balletje” means ball, like a golf ball or football. Here it means that cancer treatment is like a fast rolling ball, so that’s the metaphor”.

From the interview transcriptions, there are many broken and incomplete sentences, e.g. from Doc-0, Original sentence: “En, maar met [naam arts] nu, dat gaat een stuk beter en we hebben sinds dat wij, dat hij in [plaatsnaam] is gaan kuren, heb je zo’n oncologieverpleegkundige en daar kan je wel je vragen aan kwijt en zij... (*But now with [name of doctor], it goes a lot better and since we’ve been at [place name] to cure, you have an oncology nurse there where you can put your questions and she...*)”. The metaphors we collected using two data sets are also different in the domain words that people used to map to (the vehicle).

From our experimental outputs, different LLMs can extract and suggest different metaphors from

the same data we used even though the overlaps happen sometimes. So, it is useful to have a *collective of LLMs* to carry out such a task from our findings.

We have collected metaphors that patients can use not only in the cancer treatment process but also for hospitals, clinicians, and families, which can be very useful in a diverse way. From a clinical perspective, the extracted metaphors provide insight into how patients conceptualise illness, which can help clinicians tailor communication strategies. This aligns with prior work showing that metaphor-sensitive communication improves patient understanding and emotional support.

6. Conclusions and Future Work

In this work, we explored different kinds of prompting techniques for Dutch metaphor identification and extraction in cancer domain using two data sets including interview data transcription (oral language) and forum data (written language). Our findings provide practical guidance for prompt design in metaphor extraction tasks and demonstrate the feasibility of building metaphor datasets using local LLMs. We collected 130 metaphors (65 each) from two data sets after human verification and mapped some of the metaphors to the original English Metaphor Menu. Translation of metaphors between languages is challenging due to cultural and conceptual differences. However, the collected metaphors cover all the categories that the original English metaphor menu has, with extended new types, and are further grouped into word, phrase, and sentence levels.

This collection of Dutch metaphors can be applied to support cancer patient care by integrating them into the care-path. For the current study, we only used 100 posts of the forum data we have. For future work, we will explore more data from the blog post for extended experiments. We will look into some of the MWEs identified by LLMs, since they always overlap with idioms and metaphors (Han et al. (2025)). We also plan to explore the interpretability and explainability LLMs, incorporating formal annotation protocols, and evaluating additional models.

7. Limitations

This study has several limitations. First, the dataset size remains relatively small, particularly for the interview data. While sufficient for exploratory analysis, larger datasets would enable more robust quantitative evaluation. Although formal inter-annotator agreement metrics such as Cohen's Kappa were not computed, all annotations were independently performed and disagreements were

resolved through discussion to reach consensus. Third, the prompting strategies were evaluated primarily on local LLMs. Results may differ for larger commercial models. Fourth, distinguishing metaphors from idioms and conventionalised expressions remains challenging, both for LLMs and human annotators.

8. Ethical Statement

For both storytelling data (interviews) and kanker.nl data, we have gained ethical approvals within the research project to use. For the forum data, we used the secure version of OpenAI GPT5 established by the agreement between the University and OpenAI on anonymised and paraphrased data.

9. Acknowledgements

We thank the reviewers for valuable comments on improving our paper. We thank Ida Korfage, Sheila Payne, Judith Spek, for the feedback on the earlier abstract of this work. We thank Richele and Iris Su Yi Tamminga from USC for their valuable discussion on the extracted metaphors. Funded by the European Union under Horizon Europe Work Programme 101057332. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The UK team are funded under the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number: 10041120.

10. Bibliographical References

- Lynda Appleton and Maria Flynn. 2014. Searching for the new normal: Exploring the role of language and metaphors in becoming a cancer survivor. *European Journal of Oncology Nursing*, 18(4):378–384.
- Silke Creten, Priscilla Heynderickx, and Sylvain Dieltjens. 2022. The stigma toward dementia on twitter: a sentiment analysis of dutch language tweets. *Journal of Health Communication*, 27(10):697–705.
- Sabine De Knop and Julien Perrez. 2014. Conceptual metaphors as a tool for the efficient teaching of dutch and german posture verbs. *Review of Cognitive Linguistics. Published under the auspices of the Spanish Cognitive Linguistics Association*, 12(1):1–29.

- Jana Declercq, Lotte van Poppel, and Thomas Velvis. 2023. Machines, journeys, prisons and yo-yos: Metaphors of pain, illness and medicine in consultations with chronic pain patients. *Social Science & Medicine*, 330:116043.
- Zsófia Demjén and Elena Semino. 2020. Communicating nuanced results in language consultancy: The case of cancer and the violence metaphor. In *Professional communication: Consultancy, advocacy, activism*, pages 191–210. Springer.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. [A survey on in-context learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.
- Matteo Fuoli, Weihang Huang, Jeannette Littlemore, Sarah Turner, and Ellen Wilding. 2026. Metaphor identification using large language models: A comparison of rag, prompt engineering, and fine-tuning. *Applied Corpus Linguistics*, page 100204.
- Raymond W Gibbs Jr and Heather Franks. 2002. Embodied metaphor in women’s narratives about their experiences with cancer. *Health communication*, 14(2):139–165.
- Ingeborg PM Griffioen, Judith AC Rietjens, Marijke Melles, Dirk Snelders, Marjolein YV Homs, Casper H van Eijck, and Anne M Stiggelbout. 2021. The bigger picture of shared decision making: a service design perspective using the care path of locally advanced pancreatic cancer as a case. *Cancer Medicine*, 10(17):5907–5916.
- Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.
- Kristine J Harrington. 2012. The use of metaphor in discourse about cancer: a review of the literature. *Clinical journal of oncology nursing*, 16(4).
- Marij A Hillen, Maria Grazia Rossi, Aranka Akkermans, Leonie AL Baatenburg de Jong, Jolanda HM Dobbe, Tanja Henkel, Vicky Lehmann, Susan J Oudbier, Tanja J de Rijke, Ellen MA Smets, et al. 2025. Clinicians’ use of metaphoric language in conversations with families of critically ill patients in the intensive care unit. *Patient Education and Counseling*, page 108813.
- Charlotte Hommerberg, Anna W Gustafsson, and Anna Sandgren. 2020. Battle, journey, imprisonment and burden: Patterns of metaphor use in blogs about living with advanced cancer. *BMC Palliative Care*, 19(1):59.
- George Lakoff. 1993. The contemporary theory of metaphor.
- Yufeng Liu, Elena Semino, Judith Rietjens, and Sheila Payne. 2024. Cancer experience in metaphors: patients, carers, professionals, students—a scoping review. *BMJ Supportive & Palliative Care*, 14(e3):e2366–e2376.
- Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. [The prompt makes the person\(a\): A systematic evaluation of sociodemographic persona prompting for large language models](#).
- Dalia Magaña and Teenie Matlock. 2018. How spanish speakers use metaphor to describe their experiences with cancer. *Discourse & communication*, 12(6):627–644.
- James Chike Nwankwo. 2024. Metaphor and therapeutic potential in cancer discourse: A comparative case study of american and nigerian self-help books. *Communication & Medicine*, 19(3):271–283.
- Trijntje Pasma. 2011. Metaphor and register variation: The personalization of dutch news discourse.
- Tryntje Pasma. 2012. Metaphor identification in dutch discourse. In *Metaphor in use*, pages 69–84. John Benjamins Publishing Company.
- Tryntje Pasma. 2019. Chapter 5. linguistic metaphor identification in dutch. In *Metaphor Identification in Multiple Languages: MIPVU around the world*, pages 91–112. John Benjamins Publishing Company.
- Richard T Penson, Lidia Schapira, Kristy J Daniels, Bruce A Chabner, and Thomas J Lynch Jr. 2004. Cancer as metaphor. *The oncologist*, 9(6):708–716.
- Valeria A Pfeifer, Karen L Weihs, and Vicky T Lai. 2024. Narratives about cancer: What metaphors can tell us about depressive symptoms in breast cancer patients. *Health Communication*, 39(9):1888–1898.
- Libo Ren, Yee Man Ng, and Lifeng Han. 2025. Malei at multiclinsum: Summarisation of clinical documents using perspective-aware iterative self-prompting with llms. *arXiv preprint arXiv:2509.07622*.
- Pablo Romero, Libo Ren, Lifeng Han, and Goran Nenadic. 2025. [The Manchester bees at Per-AnsSumm 2025: Iterative self-prompting with](#)

claude and o1 for perspective-aware health-care answer summarisation. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 340–348, Albuquerque, New Mexico. Association for Computational Linguistics.

Abulhair Saparov and He He. 2023. [Language models are greedy reasoners: A systematic formal analysis of chain-of-thought](#). In *The Eleventh International Conference on Learning Representations*.

John Schuster, Erik Beune, and Karien Stronks. 2011. Metaphorical constructions of hypertension among three ethnic groups in the netherlands. *Ethnicity & health*, 16(6):583–600.

Elena Semino, Zsófia Demjén, Jane Demmen, Veronika Koller, Sheila Payne, Andrew Hardie, and Paul Rayson. 2017a. The online use of violence and journey metaphors by patients with cancer, as compared with health professionals: a mixed methods study. *BMJ supportive & palliative care*, 7(1):60–66.

Elena Semino, Zsófia Demjén, Andrew Hardie, Sheila Payne, and Paul Rayson. 2017b. *Metaphor, cancer and the end of life: A corpus-based study*. Routledge.

11. Language Resource References

Han, Lifeng and Mohamed, Najet Hadj and Rassem, Malak and Jones, Gareth JF and Smeaton, Alan F and Nenadic, Goran. 2025. [Towards a resource for multilingual lexicons: an MT assisted and human-in-the-loop multilingual parallel corpus with multi-word expression annotation](#). GitHub Open-source. AlphaMWE 2.0.

Metaphor Menu, Lancaster University. 2025. [A Metaphor Menu for people living with Cancer](#). Department of Linguistics and English Language, Lancaster University. Lancaster University, 1.0.

A. Reproducibility

To support reproducibility, we release:

- all prompts used in the study,
- model configurations,
- extraction framework code,
- paraphrased and synthetic metaphor examples.

Due to privacy restrictions, original patient texts cannot be shared. The statistics of HealthQuote.nl is shown in Table 3.

Dataset	Size
Interview documents	13
Blog posts analysed	100
Validated metaphors (interview)	65
Validated metaphors (blog)	65
Total validated metaphors	130

Table 3: HealthQuote.NL dataset statistics

B. Dutch Alignment to English Metaphor Menu

To examine cross-lingual alignment between Dutch metaphors and the English Metaphor Menu, we applied GPT-5 to the forum data and asked it to map each extracted Dutch metaphor to one of the 17 conceptual categories defined in the original English resource. Figures 4 and 5 present paraphrased examples together with the corresponding Dutch metaphor, the mapped Metaphor Menu category ("map"), the model's confidence score, and an explanation of the alignment.

The results show that many Dutch metaphors used by cancer patients correspond closely to the conceptual domains identified in the English Metaphor Menu, including battle (e.g., "vechten als een leeuw"), journey and progress (e.g., "stappen gezet"), unwanted guest or lingering presence (e.g., references to scars or body parts as companions), and emotional turbulence (e.g., "emotionele balletje dat stuitert"). The confidence scores reflect the model's estimation of semantic similarity between the Dutch metaphor and the English conceptual category, with higher scores indicating stronger alignment.

These findings suggest that the conceptual metaphor structures underlying cancer experiences are largely shared across languages, while still allowing for culturally specific linguistic realizations. This alignment supports the feasibility of extending the Metaphor Menu framework to Dutch and provides further validation for the metaphors included in the HealthQuote.NL dataset.

Blog text (paraphrased)	Metaphor (in Dutch)	map	Confidence Score	Explanation
Daar zit je dan... in de wachtkamer bij dokter De Leeuw. Goede naam, denk ik nog, hopelijk vecht ze ook als een leeuw. Mijn man zit zenuwachtig zijn handen samen te knijpen. O, wat had ik hem dit graag bespaard. ... De dokter vertelt waar ik een groot litteken zal krijgen. Ik maak een stom grapje: ik ben toch al gelukkig getrouwd. ... Mijn gedachten dwalen af. Lymfeklieren, overall in je lichaam... Ineens raakt het me volledig, heel kort maar glashelder. De brok in mijn keel voel ik, maar ik druk de tranen weg. Nee, dat sta ik mezelf niet toe. De kinderen... denk aan de kinderen. We gaan gewoon voor goede uitslagen. Ik zal vechten als een leeuw. Als je wilt, kan ik er ook een meer literaire, bijna poëtische versie van maken, die de spanning en emoties nóg sterker benadrukt, zonder dat het te zwaar wordt. Wil je dat?	Hopelijk vecht ze als een leeuw.	7	0.95	Explicit simile of fighting like a lion — directly expresses a combat/fight frame, matching item 7 (I don't intend to give up; I want to fight).
	een groot litteken	13	0.72	A scar functions as a lasting trace/after-effect of illness/treatment; thematically closest to item 13 (cancer as an unwelcome lodger that leaves a mess / lasting memory/residue).
	Ineens raakt het me volledig, heel kort maar glashelder.	2	0.85	This describes sudden clarity and heightened perception; maps best to item 2 (the illness makes you look up and take notice of the scenery / increased appreciation/awareness)
	Ik zal vechten als een leeuw.	7	0.99	Straightforward, first-person battle metaphor — very strong direct fit with item 7 (determined combat).
Na een week wachten mochten we eindelijk de uitslag van het biopt van een vlekje in mijn mond horen. Volgens Dokter Google zou het zomaar een onschuldige "amalgamtattoo" kunnen zijn, maar stiekem bleef in mijn achterhoofd het idee rondspoken dat het ook een melanoom kon zijn. Dankzij hem konden we snel weer naar huis. Nu op naar woensdag: het intakegesprek over de immunotherapie, waarvan de eerste behandeling donderdag al op het programma staat.	Volgens Dokter Google kon het een onschuldige 'amalgamtattoo' zijn, ... melanoom kon zijn.	1	0.65	The 'haunting in the back of my mind' gives a feeling of fear and uncertainty that is temporary, which fits item 1 (scary fairground ride)
	konden we snel weer naar huis	5	0.25	Returning home after reassurance suggests social/support element; loosely aligns with item 5 (the rocks in our path are easier with others who help you). Low confidence because the phrase is largely literal (going home) and not a clear metaphor of shared burden.
	Op naar woensdag, het intakegesprek over de immunotherapie...	3	0.30	Phrase implying forward motion toward a next step; loosely maps to a journey metaphor. Best (but weak) fit is item 3 (journeys vary in length; what matters is how we walk the journey) because it signals continuing along a process. Low confidence because the expression is mostly temporal/planning rather than an explicit journey metaphor.

Figure 4: Metaphor identification using GPT5 on forum data (paraphrased) with model confidence scores: part-1 ('map' column refers to the value in original English Metaphor menu)

Blog text (paraphrased)	Metaphor (in Dutch)	map	Conf. Score	Explanation
... Ik ben de afgelopen weken druk geweest met alles rondom mijn reconstructie, gesprekken met mijn psycholoog, rebalancing en andere dingen om alles een plek te geven, en ook om mijn werkuren langzaam weer op te bouwen. Voor mijn gevoel heb ik behoorlijk wat stappen gezet, en 70% van de tijd gaat het eigenlijk best goed. ... Dat blijft nog steeds het meest onwerkelijke van alles: zo heb je er twee, en ineens nog maar één. ... Toen ik gisteren naar de fysio reed en langs het ziekenhuis kwam, kreeg ik ineens flashbacks van de rit naar het ziekenhuis voor de amputatie, inclusief de angst en het verdriet die ik toen voelde, en de terugweg na de operatie, toen mijn plezier in het leven ineens leek te verdwijnen tegelijk met mijn borst. Het rauwe verdriet kwam ineens terug. ... Het helpt ook niet dat ik vaak droom dat ik op de operatietafel lig, wachtend op de chirurg, of dat ik me probeer te verstoppen. Sinds de amputatie droom ik regelmatig over geopereerd worden. Het zijn geen fijne dromen. Soms vind ik het gewoon zo'n raar idee. Mijn borst heeft me vanaf mijn puberteit vergezeld, een groot deel van mijn leven, en binnen een uur haalt een chirurg hem weg. ... Het onwerkelijke gevoel blijft ook. Soms lijkt het alsof mijn brein achterloopt, alsof het denkt: ... Ik was er immers bij, maar soms voelt het alsof het niet echt mij is overkomen, of alsof ik het niet volledig kan bevatten. Ik zie een lege plek, en fysiek weet ik dat ik een operatie heb gehad, maar het voelt soms alsof ik nog in een droom of nachtmerrie zit waaruit ik moet ontwaken. ... De plastisch chirurg legde uit dat dit normaal is: het brein communiceert met alle lichaamsdelen en ontvangt signalen terug, maar niet van de delen die weg zijn, waardoor het brein concludeert dat er iets niet klopt. Toch kijk ik uit naar het moment dat het emotionele "balletje" wat minder hard stuitert.	om alles een plek te geven	3	0.65	Putting things in place = processing and moving along; maps reasonably to type 3 (the journey and how we walk it — making steps and arranging things emotionally).
	wat stappen gezet	3	0.95	'Steps' is a clear journey/process metaphor — strong match to type 3.
	zo heb je er twee, en ineens nog maar één	1	0.75	The suddenness/disorientation evokes abrupt jolting change—loosely like type 1 (scary ride/abrupt experience).
	mijn plezier in het leven ineens leek te verdwijnen tegelijk met mijn borst	9	0.45	Equating loss of pleasure with loss of the breast suggests ongoing relationship/ebb-and-flow of experience; type 9 (cancer as relationship with shifting control) is a loose interpretive fit.
	droom dat ik op de operatietafel lig	14	0.50	Dreams and hiding suggest alienation from the body and invasive experience; relates to type 14 (initial alien-invasion/feeling not connected to changed body)
	mijn borst heeft me ... vergezeld	13	0.75	Breast as a companion/personified presence resonates with type 13 (cancer/part as an unwelcome/lingering lodger/stranger) — emphasizes loss and relationship to the body-part.
	binnen een uur haalt een chirurg hem weg	13	0.80	Reduction of a breast to 'a piece of tissue' / discarded object implies dispossession and stranger/other framing — aligns with type 13 (unwanted lodger/alien presence and loss).
	alsof mijn brein achterloopt	14	0.80	The 'brain lagging' image captures disconnection between mind and changed body — a strong match to type 14 (feeling alienated/not connected to bodily change).
	ik zie een lege plek	6	0.60	An ongoing felt absence (empty place) causing discomfort resonates with type 6 (persistent irritant/limitation after cancer — stone in the shoe metaphor for ongoing ache/discomfort).
	het brein communiceert ... maar niet van de delen die weg zijn, waardoor het brein concludeert dat er iets niet klopt	14	0.70	This physiological personification (brain-body dialogue) highlights disconnection and 'not right' sensation — thematically fits type 14 (initial alienation / not integrated).
	dat emotionele 'balletje' wat minder hard stuitert	1	0.92	Same bouncing-ball image; explicitly about calming emotional oscillation — strong match to type 1.

Figure 5: Metaphor identification using GPT5 on forum data (paraphrased) with model confidence scores: part-2 ('map' column refers to the value in original English Metaphor menu)

Compressed Representations of Patient Records: A Comparative Study of Template-Based and LLM-Based Methods for Clinical Data Summarization and Visualization

Andreas Stöckl, Oliver Krauss, Sophie Bauernfeind

Digital Media Lab, Digital Media Lab, Research & Development

University of Applied Sciences Upper Austria

Hagenberg, Austria

{andreas.stoeckl, oliver.krauss, sophie.bauernfeind}@fh-hagenberg.at

Abstract

Electronic Health Records (EHRs) contain comprehensive patient information that is often voluminous and challenging to review efficiently. This paper presents a systematic evaluation of multiple methods for compressing patient records into standardized, comparable formats. Four compression approaches are implemented and compared: two template-based methods (structured extraction, extractive key-phrase) and two LLM-based methods (LLM, and hybrid LLM with 8 different models). Using a synthetic cohort of 75 patient records generated with realistic clinical patterns, each method is evaluated on information preservation (diagnosis, medication, allergy, lab value recall, and vital accuracy), compression efficiency, and output quality. Across methods, diagnosis recall ranged from 0.637 to 1.000, with medication and allergy recall consistently exceeding 0.880. In the test setup, the template-based approach yielded the highest compression ratio (7.6×), while the hybrid methods provided the most balanced trade-off between compression and clinical utility. These results suggest that combining structured extraction with LLM-generated summaries can be an effective strategy for scenarios requiring both compact representations and contextual clinical information.

Keywords: Electronic Health Records, Clinical Text Summarization, Large Language Models, Data Compression, Clinical NLP, Infographics

1. Introduction

The proliferation of Electronic Health Records (EHRs) has created both opportunities and challenges for healthcare delivery. While comprehensive digital records enable better continuity of care, the sheer volume of clinical documentation often exceeds what clinicians can efficiently review (Shickel et al., 2018). A typical patient encounter may require reviewing hundreds of pages of prior records, laboratory results, medication lists, and clinical notes (Pivovarov and Elhadad, 2015).

This information overload has motivated significant research into clinical text summarization and EHR compression. Recent advances in Large Language Models (LLMs) have shown promising results in generating coherent clinical summaries (Van Veen et al., 2024; Schilder et al., 2025; Schoonbeek et al., 2025). However, questions remain about the optimal approach for different clinical use cases, particularly when the goal is to enable rapid comparison across multiple patients.

This work investigates methods for generating compressed representations of patient records that retain essential clinical information while supporting efficient review. The contributions include:

1. A systematic comparison of four compression methods spanning template-based and LLM-based approaches with different models.

2. A synthetic patient dataset with realistic clinical patterns for reproducible evaluation.
3. Novel evaluation metrics focused on information preservation and comparability.
4. An open-source implementation with infographic visualization tools.

2. Related Work

2.1. Clinical Text Summarization

Clinical text summarization has evolved significantly with the advent of deep learning. Early approaches relied on extractive methods that selected key sentences from source documents (Pivovarov and Elhadad, 2015). More recent work has explored abstractive summarization using transformer architectures (Shickel et al., 2018).

Van Veen et al. (Van Veen et al., 2024) demonstrated that adapted LLMs can outperform medical experts in clinical text summarization tasks. Their study applied adaptation methods to eight LLMs across four clinical summarization tasks, finding that in 45% of cases LLM summaries were equivalent to expert summaries, and in 36% of cases they were superior.

A recent scoping review (Schilder et al., 2025) analyzed 30 studies on LLM-based clinical text

summarization, identifying key limitations including narrow research focus (57% on radiology reports), heavy reliance on the MIMIC dataset (50% of studies), and insufficient clinical impact evaluation.

2.2. EHR Representation Learning

Deep representation learning for EHRs has been extensively studied (Si et al., 2021). The Graph Convolutional Transformer (Choi et al., 2020) demonstrated the value of learning graphical structures from EHR data. BioBERT (Lee et al., 2020) and Clinical BERT (Alsentzer et al., 2019) have shown strong performance on biomedical NLP tasks.

2.3. Large Language Models in Healthcare

GPT-4 (OpenAI, 2023) and similar LLMs have demonstrated remarkable capabilities in medical knowledge tasks (Singhal et al., 2023). Tang et al. (Tang et al., 2023) evaluated LLMs on medical evidence summarization, finding promising but inconsistent results. Concerns remain about hallucination and factual accuracy in clinical contexts (Wornow et al., 2023). MedGemma models were built upon Gemma 3, with optimization for medical domains, and has been shown to outperform other generative models of comparable size on clinically oriented tasks. (Sellersgren et al., 2025)

2.4. Synthetic EHR Generation

Synthea (Walonoski et al., 2018) provides an open-source framework for generating realistic synthetic patient records. The synthetic data approach addresses privacy concerns while enabling reproducible research. Synthea generates complete patient histories conformant with FHIR standards, including diagnoses, medications, procedures, and laboratory values.

3. Methods

3.1. Dataset Generation

A synthetic cohort of 75 patient records was generated using a custom generator inspired by Synthea (Walonoski et al., 2018). Each record includes:

- Demographics (name, age, gender, MRN)
- Diagnoses with ICD-10 codes (1-7 per patient)
- Medications with dosing (1-12 per patient)
- Allergies with reaction severity
- Vital signs (BP, HR, RR, SpO2, weight, ...)

- Laboratory results with reference ranges and abnormal flags
- Social and family history
- Clinical notes (1,500-2,500 characters)

Patient complexity was varied across three levels (low: 29 patients, moderate: 33 patients, high: 13 patients), affecting the number of diagnoses, medications, and abnormal findings.

3.2. Example Patient Record

Listing 1 shows an excerpt from a synthetic patient record. The full clinical note spans approximately 1,874 characters and includes structured sections for chief complaint, history, medications, allergies, vital signs, physical examination, laboratory results, and assessment.

CHIEF COMPLAINT: COPD exacerbation

HISTORY OF PRESENT ILLNESS:

Patient presents COPD exacerbation. Patient reports compliance with current medications. Ongoing management of Urinary tract infection.

PAST MEDICAL HISTORY:

- Urinary tract infection (N39.0)

MEDICATIONS:

- Aspirin 81mg daily
- Sertraline 50mg daily
- Lisinopril 10mg daily

ALLERGIES:

- No known allergies

VITAL SIGNS:

- Blood Pressure: 119/72 mmHg
- Heart Rate: 65 bpm
- Oxygen Saturation: 100%
- Weight: 111.9 kg, Height: 173 cm
- BMI: 37.4

LABORATORY RESULTS:

- WBC: 8.41 K/uL (4.5-11.0)
- Hemoglobin: 15.94 g/dL (12.0-17.5)
- Glucose: 92.76 mg/dL (70-100)
- eGFR: 49.0 mL/min (90-120) [L]
[...]

Listing 1: Excerpt from original patient record (Patient MRN000001, 56-year-old male). Full record: 1,874 characters.

3.3. Compression Methods

In total, four distinct compression methods were evaluated. The LLM-based approach and the hy-

brid LLM approach each incorporated eight different models, resulting in 18 individual compression tests across all four methods:

3.3.1. Template-Based Extraction

A rule-based approach that extracts key information using predefined templates without natural language generation. Key elements (diagnoses, medications, vitals, abnormal labs) are directly extracted and formatted into a standardized structure.

3.3.2. Extractive Key-Phrase

Uses TF-IDF-inspired scoring to identify and extract the most clinically significant phrases from the patient record. Medical terms are weighted by category (diagnoses: 3.0, allergies: 3.0, medications: 2.5, abnormal labs: 2.0).

3.3.3. LLM

Uses different LLM models to generate structured clinical summaries. The model receives formatted patient data and returns JSON-structured output including primary diagnoses, vitals summary, risk factors, active allergies, active medications, critical labs, clinical summary and care priorities. The evaluation included both cloud-based and self-hosted model deployments:

- GPT-4o-mini (cloud)
- GPT-4o (cloud)
- google_medgemma-4b-it-Q6_K_L (local)
- google_medgemma-27b-it-Q6_K_L (local)
- gemma-3-12b-it-Q4_K_M (local)
- lfm2.5-thinking (local)
- Meta-Llama-3.1-8B-Instruct-Q6_K_L (local)
- nemotron-3-nano (local)

Prompt: No prompt engineering was performed to avoid introducing potential bias in favor of specific models. Instead, a single concise and clear instruction set was developed that consistently produced valid outputs across all evaluated systems.

3.3.4. Hybrid-LLM

Combines template-based structured extraction, described in subsection 3.3.1 with LLM-generated clinical narratives. Structured data (vitals, labs, medications) are extracted algorithmically, while the LLM generates a comparative-friendly clinical summary. The models used in the hybrid LLM evaluation were identical to those utilized in the standard LLM approach.

You are analyzing a patient record for
→ clinical summarization.
Create a comprehensive yet concise
→ clinical summary optimized for:
1. Quick clinical review
2. Comparison with other patients
3. Identifying key risks and priorities

Patient Record:
[...]

Provide your analysis in this exact JSON
→ structure:

```
{
  "primary_diagnoses": ["List
    → diagnoses in order of clinical
    → significance"],
  "vitals": "List important vitals as
    → short plain text",
  "active_allergies": ["List important
    → known allergies"],
  "active_medications": ["List active
    → medications as plain text
    → including name, dose and
    → frequency"],
  "risk_factors": ["Identified risk
    → factors for adverse outcomes"],
  "critical_labs": ["List abnormal
    → labs as plain text including
    → values and clinical
    → interpretation"],
  "care_priorities": ["Top 3 care
    → priorities in order of
    → urgency"],
  "clinical_summary": "A 2-3 sentence
    → clinical overview covering: main
    → conditions, current status, and
    → key concerns"
}
```

Respond ONLY with valid JSON.

Listing 2: Prompt structure for LLM compression, with [...] being the text of the patient shown in [Listing 1](#).

Prompt: The prompt focuses on the overall summary, with the already extracted information to reduce the input token.

3.4. Output Format

All methods produce a standardized Compressed-PatientRecord containing:

- Patient ID, name, age, gender
- Primary diagnoses (list)
- Active medications (list with dosing)
- Allergies

Based on this patient summary, generate
 ↳ a 2-3 sentence clinical overview
 ↳ optimized for comparing with other
 ↳ patients:

```
Patient: {template_result.name},
↳ {template_result.age}y
↳ {template_result.gender}
Diagnoses: {'', '.join(template_result.p
↳ rimary_diagnoses)}
Vitals: {template_result.vital_summary}
Abnormal Labs: {'', '.join(template_resu
↳ lt.critical_labs)}
Risk Factors: {'', '.join(template_resul
↳ t.risk_factors)}
```

```
Provide your analysis in this exact JSON
↳ structure:
{{
  "clinical_summary": "2-3 sentence
↳ clinical overview optimized for
↳ comparing with other patients."
}}
```

```
Focus on: disease severity, control
↳ status, and key concerns. Be concise
↳ and just return the json.
```

Listing 3: Prompt to create summary for Hybrid Compression strategy using extracted template information.

- Vital signs summary (one-line format, including values of BP, HR, SpO2, weight)
- Critical/abnormal laboratory values
- Identified risk factors
- Clinical summary narrative (LLM methods only)

3.5. Evaluation Metrics

The methods were evaluated on multiple dimensions:

Information Preservation:

- *Diagnosis Recall*: Proportion of original diagnoses present in compressed output
- *Medication Recall*: Proportion of original medications preserved
- *Allergy Recall*: Proportion of allergies correctly captured
- *Lab Abnormal Recall*: Proportion of abnormal lab values identified
- *Vital Accuracy*: Correctness of vital sign representation

Compression Metrics:

- *Compression Ratio*: Original length / Compressed length
- *Field Completeness*: Proportion of expected fields populated (without clinical summary)

Quality Score: A composite metric combining preservation (40%), field completeness (25%), compression efficiency (15%), and presence of clinical summary (20%).

3.6. Visualization

An infographic visualization system was developed to render compressed patient records as interactive HTML dashboards. The visualization includes:

- Patient cards with color-coded risk indicators
- Comparison tables for cohort-level review
- Aggregate statistics (mean age, diagnoses, medications)
- Highlighted abnormal values and risk factors

Figure 1 shows a screenshot of the patient cohort dashboard. The header displays aggregate statistics (75 patients, average age 57, 2.9 diagnoses per patient, 3.7 medications per patient). Below, individual patient cards present compressed information with visual risk indicators (color-coded badges showing LOW/MEDIUM/HIGH risk). Each card displays vital signs with color-coded borders for abnormal values, diagnosis chips colored by severity, medication lists, and critical laboratory findings.

3.7. Compression Output Examples

Listing 4 shows the same patient as in Listing 1 with compressed record using three different methods, demonstrating the trade-off between compression ratio and narrative richness.

4. Results

4.1. Information Preservation

Table 1 presents the comparative results across all methods. The diagnosis recall rate varied between 0.637 (LLM-lfm2.5) and 1.000 (Extractive Key-Phrase & Template-Based), while the medication recall rate and allergy recall rate all had values above 0.880.

Lab abnormal recall varied between 0.346 (LLM-lfm2.5ing) and 1.000 (Extractive Key-Phrase). The slight reduction in LLM methods occurred when models reformulated lab findings rather than just list the exact values and some also misses flagged

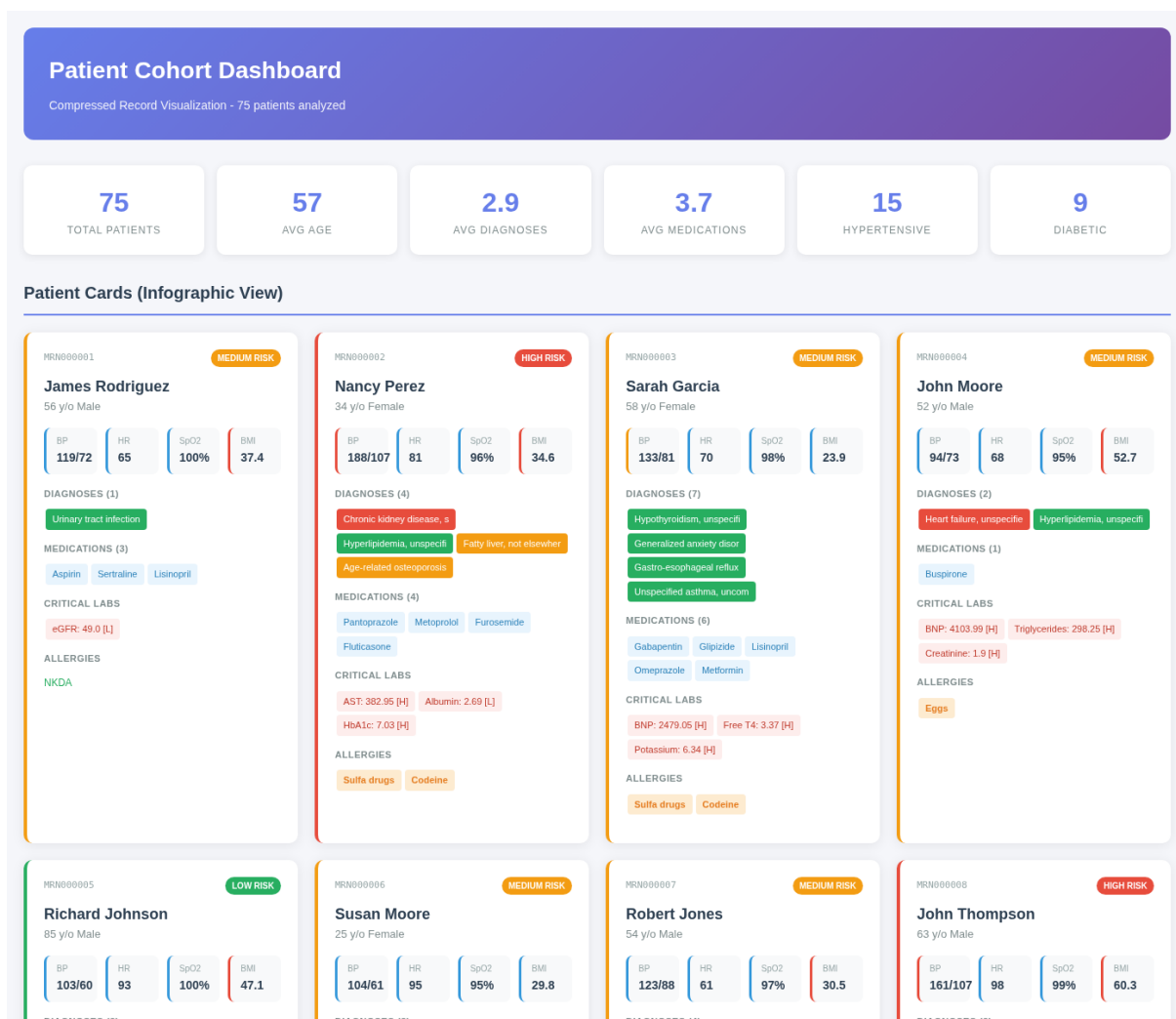


Figure 1: Patient Cohort Dashboard visualization. The header shows aggregate cohort statistics. Patient cards display compressed records with color-coded risk levels, base on the complexity (green=low, orange=medium, red=high), vital signs, diagnoses, medications, and critical lab values. This infographic format enables rapid visual comparison across patients.

lab values. As for the template-based compression is concerned, only the first 5 values were taken to achieve compression.

Vital sign accuracy varied between 0.600 (Extractive Key-Phrase) and 0.997 (LLM-gemma-3-12b).

As expected, there was hardly any difference in the quality score for the hybrid variant, as the difference between the methods is only apparent in one field, the summary, which is discussed in 4.4. The improved score of the lfm2.5 models can be explained by the higher compression rate.

4.2. Compression Efficiency

Template-Based extraction achieved the highest compression ratio (7.6x), reducing average clinical notes from approximately 1,874 characters to 247 characters. This aggressive compression comes at the cost of excluding clinical narratives. The Extractive Key-Phrase method has a lower compression

ratio of 4.1.

In this setup, LLM methods produced the lowest compression ratios, ranging from 3.2 to 3.8, due to the inclusion of generated clinical summaries. The Hybrid-LLM method balanced compression (4.5x) with narrative quality.

4.3. Quality Comparison

The Hybrid-LLM method achieved the highest overall quality score (0.88), followed by the LLM approach using GPT-4o (0.878), the Key-Phrase method (0.868), and the remaining LLM models (0.783–0.868). In contrast, the Template-Based method yielded the lowest overall quality score (0.729).

The quality improvement from LLM methods primarily derives from the clinical summary narratives, which provide interpretive context not available in template-only approaches.

Template-Based Extraction (257 chars, 7.3x compression):

[MRN000001] James Rodriguez, 56y Male
Dx: Urinary tract infection
Meds: Aspirin, Sertraline, Lisinopril
Allergies: NKDA
VS: BP 119/72, HR 65, SpO2 100%
Labs: eGFR: 49.0 mL/min [L]
Risks: Obesity (BMI 37.4), Active smoker, Reduced kidney function (eGFR 49.0)

Hybrid-medgemma-27b (467 chars, 4.0x compression):

[MRN000001] James Rodriguez, 56y Male
Dx: Urinary tract infection
Meds: Aspirin, Sertraline, Lisinopril
Allergies: NKDA
VS: BP 119/72, HR 65, SpO2 100%
Labs: eGFR: 49.0 mL/min [L]
Risks: Obesity (BMI 37.4), Active smoker, Reduced kidney function (eGFR 49.0)
Summary: 56-year-old male presenting with a urinary tract infection. He has significant risk factors including obesity (BMI 37.4), active smoking, and reduced kidney function (eGFR 49.0 mL/min), which may impact treatment and recovery.

LLM-gpt-4o-mini (563 chars, 3.3x compression):

[MRN000001] James Rodriguez, 56y Male
Dx: Urinary tract infection (N39.0)
Meds: Aspirin, Sertraline, Lisinopril
Allergies: NKDA
VS: BP: 119/72 mmHg, HR: 65 bpm, SpO2: 100%, Weight: 111.9 kg
Labs: eGFR: 49.0 mL/min [L] - indicates potential renal impairment
Risks: Current smoker (0.5 pack/day), Family history of stroke and colon cancer, Elevated weight
Summary: James Rodriguez is a 56-year-old male diagnosed with a urinary tract infection. Current management includes medications for cardiovascular health and mental health, but significant risk factors include his smoking status and family history of chronic diseases.

Listing 4: Comparison of compression outputs for Patient MRN000001 using three methods. Template-based achieves highest compression; LLM methods provide clinical interpretation.

Method	Dx Recall	Med Recall	All. Recall	Lab Recall	Vital Acc	Comp. Ratio	Field Comp.	Preserv.	Quality
Template-Based	1.000	1.000	1.000	0.943	0.800	7.6x	0.940	0.949	0.729
Extractive Key-Phrase	1.000	1.000	1.000	1.000	0.600	4.1x	0.953	0.920	0.868
LLM-Meta-Llama	0.806	0.998	0.973	0.520	0.837	3.2x	0.967	0.827	0.820
LLM-gemma-3-12b	0.821	0.987	0.987	0.916	0.976	3.5x	0.953	0.937	0.865
LLM-medgemma-27b	0.835	0.972	1.000	0.939	0.997	3.2x	0.967	0.949	0.868
LLM-medgemma-4b	0.837	1.000	1.000	0.728	0.859	3.4x	0.967	0.885	0.847
LLM-llm2.5	0.637	0.970	0.880	0.346	0.808	3.8x	0.940	0.728	0.783
LLM-nemotron-3	0.831	0.887	0.907	0.747	0.931	3.4x	0.962	0.860	0.836
LLM-gpt-4o-mini	0.967	0.995	1.000	0.917	1.000	3.0x	0.971	0.976	0.879
LLM-gpt-4o	0.906	1.000	0.987	0.938	0.992	3.0x	0.962	0.965	0.872
Hybrid-Meta-Llama	1.000	1.000	1.000	0.943	0.800	4.5x	0.940	0.949	0.882
Hybrid-gemma-3-12b	1.000	1.000	1.000	0.943	0.800	4.5x	0.940	0.949	0.882
Hybrid-medgemma-27b	1.000	1.000	1.000	0.943	0.800	4.5x	0.940	0.949	0.882
Hybrid-medgemma-4b	1.000	1.000	1.000	0.943	0.800	4.5x	0.940	0.949	0.882
Hybrid-llm2.5	1.000	1.000	1.000	0.943	0.800	4.6x	0.940	0.949	0.884
Hybrid-nemotron-3	1.000	1.000	1.000	0.943	0.800	4.5x	0.940	0.949	0.882
Hybrid-gpt-4o-mini	1.000	1.000	1.000	0.943	0.800	3.4x	0.940	0.949	0.865
Hybrid-gpt-4o	1.000	1.000	1.000	0.943	0.800	3.2x	0.940	0.949	0.863

Table 1: Comparison of Patient Record Compression Methods (average values)

Figure 2 visualizes the key metrics across all methods, highlighting the trade-off between compression ratio and quality score.

4.4. Clinical Summary Analysis

The choice of model—whether applied in the LLM or hybrid approach—has only a marginal influence

on the resulting summaries based on our manual assessment. Although the hybrid approach more consistently incorporates vital signs and their corresponding values, the overall structure and content of the outputs remain largely comparable to those produced by the full LLM approach. The following examples present summaries generated using the LLM approach:

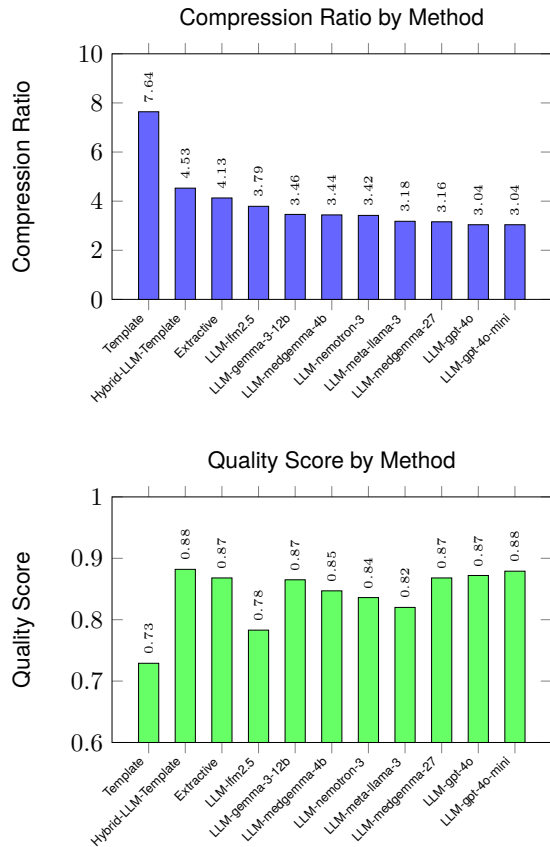


Figure 2: Infographic comparison of compression methods. Top: compression ratio; Bottom: quality score (higher is better for both). Template-Based leads in compression, Hybrid-LLM-Template and LLM-gpt-4o-mini in quality.

GPT-4o-mini produced concise 2-sentence summaries focusing on primary condition, current status and possible risks. Example: *“James Rodriguez is a 56-year-old male diagnosed with a urinary tract infection. Current management includes ... risk factors include his smoking status and family history of chronic diseases.”*

GPT-4o produced a three-sentence summary that included medication status, diagnosis, and lifestyle factors, while the family history was not always present. The final sentence highlighted recommended next steps or focus areas. Example: *“... is diagnosed with a urinary tract infection. His medical management includes ... controlled vital signs, but he is at risk due to obesity, smoking, and reduced renal function. ...”*

google_medgemma-4b-it-Q6_K_L produces a three-sentence overview: first a brief patient summary with the current diagnosis, followed by key lab values with possible interpretations, and finally treatment suggestions or important clinical considerations. Example: *“James Rodriguez, a 56-year-old male, presents ... elevated BUN/creatinine, suggesting possible early chronic kidney disease ...*

with a focus on encouraging smoking cessation.”

google_medgemma-27b-it-Q6_K_L produced a two- to three-sentence summary that included general patient information, diagnoses, and listed risk factors such as family history, while occasionally referring to or incorporating vital-sign values. Example: *“56M presents with ... current smoker with obesity and ... has Stage 3a CKD (eGFR 49). Key concerns include managing the UTI...”*

gemma-3-12b-it-Q4_K_M produced a three-sentence summary that included current diagnosis, incorporated key family history and relevant laboratory values without interpretation; and concluded with a brief treatment suggestion. Example: *“James Rodriguez is a 56-year-old male presenting ... He has a history of smoking and family history ... slightly low potassium. Management should focus on treating ...”*

llm2.5-thinking produced highly inconsistent summaries, ranging from one to three sentences; in several cases, the output contained very little or no meaningful patient information and remained overly vague. Example: *“A 56-year-old male .. with normal vitals and no allergies, managed with standard antibiotics ... due to smoking-related risks.”*

Meta-Llama-3.1-8B-Instruct-Q6_K_L produced one- to three-line summaries that included the patient’s age, gender, and a variable-detail description of medical history and current diagnosis. However, it introduced additional descriptive terms or inferred diagnoses based on the listed medications. Example: *“A 56-year-old male patient ... and a history of hypertension. He is currently taking medications for heart health and depression. The primary concern is managing the infection ...”*

nemotron-3-nano produced one- to three-line summaries that included a patient’s overview and a concise account of medical history and current diagnosis, highlighted relevant risk factors, and in some cases referenced or numerically reported vital-sign values with interpretation. Example: *“James Rodriguez is a ... his eGFR is reduced at 49 mL/min indicating early chronic kidney disease... family history of stroke ...”*

4.5. Processing Time

Non-LLM methods processed all 75 patients in under 0.1 seconds. Cloud LLM methods required 405–577 seconds (5.4–7.7 seconds per patient), while the Local hosted LLM’s are slower with 2113–6682 seconds (35–111 minutes; 28–89 seconds per patient). The local Hybrid-LLM-Template methods are an improvement of the only LLM methods and took 589–1513 seconds (7.8–20 seconds per patient) and the cloud Hybrid-LLM-Template methods took 160-204 seconds (2.1–2.72 seconds).

4.6. Tabular Patient Comparison

Table 2 demonstrates that compressed records may facilitate quicker cross-patient comparison by presenting information in a standardized format. This structure could help clinicians more easily identify patients of interest and review key clinical parameters, though further study is needed to assess its impact in real-world clinical settings. The tabular representation appears to preserve essential information while potentially reducing cognitive load, but this requires additional validation.

5. Discussion

Barretto et al. (Barretto et al., 2026) compare GPT-4o and fine-tuned Llama3 models using several heuristic metrics as well as a novel LLM-based evaluation scheme. However, their analysis is limited to this small set of models, and domain-specialized models such as MedGemma, which are particularly relevant for medical summarization tasks were not included.

The comparative analysis of template-based, key-phrase, LLM-based, and hybrid compression methods highlights several important considerations for selecting an appropriate approach depending on clinical or computational requirements. While all methods successfully condense patient records into standardized formats, their performance varies substantially in terms of information preservation, compression efficiency, interpretability, and computational cost.

5.1. Method Selection Guidelines

Our results suggest different methods are optimal for different use cases:

High-throughput screening: Template-based extraction achieves the highest compression ratio (7.6×) while maintaining perfect preservation of diagnostic, medication, and allergy information with a fast processing time. However, this approach exhibits lower performance in lab abnormal recall and vital-sign accuracy, and the absence of a clinical summary further contributes to its comparatively low overall quality score. Consequently, this method is best suited for applications requiring minimal storage or high-throughput automated processing.

Clinical decision support: Hybrid methods provide the best balance of compression and clinical utility, with quality scores exceeding 0.88. The tested LLM hybrid methods were combined with the template-based approach. A combination with the extractive key phrase approach could lead to a better quality score, but compression could suffer as a result.

Patient comparison: All methods generate standardized outputs that enable structured, table-based comparison across patients, and the accompanying infographic visualization further facilitates rapid cohort-level assessment.

5.2. LLM Considerations

While LLM-based methods provide clear advantages in producing clinical narrative summaries, they also introduce notable considerations:

Interpretability and Trustworthiness: Certain LLM generated summaries provide clinical interpretation such as inferring chronic kidney disease (CKD) from reduced eGFR values, that template-based methods cannot deliver without substantial rule-based encoding. However, as outlined in the Clinical Summary Analysis 4.4, interpretive performance varies considerably across individual LLM models. Models trained specifically on medical data, such as MedGemma, are generally better suited for tasks requiring clinically grounded reasoning. Importantly, ensuring traceability and transparency remains essential, particularly in light of regulatory requirements under the European AI Act and the Medical Devices Directive (Bednarczyk et al., 2025).

Consistency: Template methods guarantee consistent output format while LLM outputs may vary in structure despite JSON formatting instructions. Notably, the lfm model deviated from the specified schema despite clear prompting, while the other LLMs consistently adhered to the intended structure under identical conditions. Nevertheless, the task of returning a consistently structured output in an LLM poses a challenge (Lu et al., 2025; Li et al., 2024; Geng et al., 2025).

Cost and Latency: API-based LLM methods incur per-token costs and network latency. Latency is particularly high with self-hosted models. Hybrid approaches can reduce this latency while still providing a clinical summary.

Hallucination Risk: While we observed no hallucinations in our evaluation, LLM summaries should be validated for safety-critical applications (Wornow et al., 2023).

6. Conclusion

A comprehensive evaluation of four distinct compression methods for compressing patient records into standardized, comparable formats was presented. The Template-Based and Extractive Key-Phrase achieved perfect preservation of diagnoses, medications and allergies, while the LLM-based methods provided enhanced clinical narratives with a quality improvement.

Patient	Age	Gender	Primary Dx	Key Labs	Risk Factors
James Rodriguez	56	Male	Urinary tract infection	eGFR: 49.0 mL/min [L]	Obesity (BMI 37.4) Active smoker, Reduced kidney function (eGFR 49.0)
Nancy Perez	34	Female	Chronic kidney disease, stage 4 Hyperlipidemia, unspecified Fatty liver, not elsewhere classified Age-related osteoporosis without fracture	Glucose: 249.25 mg/dL [H] Creatinine: 4.66 mg/dL [H] eGFR: 26.84 mL/min [L]	Hypertension Severe hypertension Obesity (BMI 34.6) Active smoker Reduced kidney function (eGFR 26.84)
John Moore	52	Male	Heart failure, unspecified Hyperlipidemia, unspecified	BNP: 4103.99 pg/mL [H] Potassium: 5.9 mEq/L [H] LDL: 111.29 mg/dL [H]	Obesity (BMI 52.7)
Susan Moore	25	Female	Fatty liver, not elsewhere classified Alcohol dependence, uncomplicated	Glucose: 290.45 mg/dL [H] Potassium: 6.97 mEq/L [H]	Active smoker
Richard Johnson	85	Male	Unspecified asthma, uncomplicated, Gastro-esophageal reflux disease with esophagitis	Glucose: 275.53 mg/dL [H]	Obesity (BMI 47.1), Heavy alcohol use

Table 2: Tabular Comparison of Compressed Patient Records (Template-Based Extraction)

The Hybrid-LLM-Template method and the LLM-gpt-4o-mini method achieved the best overall quality (0.88). For applications requiring interpretive summaries, the Hybrid-LLM approach with the a fitting model provides clinical context while maintaining structured data fidelity.

7. Outlook

Looking ahead, an important next step will be to evaluate the proposed approaches on real-world EHR datasets. Such an assessment would provide deeper insight into practical utility in authentic clinical environments. For this test only locally hosted LLM would be used to ensure compliance with data-protection and regulatory requirements. This would limit the range of deployable models.

Moreover, real-world EHRs typically contain longer, more heterogeneous, and less standardized documentation than the synthetic records used here, which may introduce additional challenges for both compression and summarization.

Another promising direction involves engaging clinicians directly in the evaluation process. Clinical expertise is essential for assessing the usefulness, safety, and interpretability of compressed records and summaries, and clinician-involved validation would offer a more grounded understanding of how these methods support real decision-making.

8. Ethical Considerations and Limitations

The synthetic patient data was generated for research purposes only and does not represent real individuals. The study also has several limitations:

- Synthetic data:** Although realistic patterns are produced by the generator, greater variability and complexity are present in real clinical records.
- English only:** All methods were evaluated on English-language records.
- No clinical validation:** Clinical utility was not assessed in cooperation with healthcare providers or on real-world patient data.
- Prompt engineering:** In this study, a short and clear prompt was used to generate valid JSON outputs. Because prompt engineering was not applied to the individual models, they may produce weaker results as a result.

9. Bibliographical References

- Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew BA McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.
- Daphne Barretto, Matthew Jin, and Bora Oztekin. 2026. [Clinical text summarization with llm-based evaluation](#). Stanford CS224N Project.
- Lydie Bednarczyk, Daniel Reichenpfader, et al. 2025. [Scientific evidence for clinical text summarization using large language models: Scoping review](#). *Journal of Medical Internet Research*.
- Edward Choi, Zhen Xu, Yujia Li, Michael W Dusenberry, Gerardo Flores, Yuan Xue, and Andrew M Dai. 2020. Learning the graphical structure of

- electronic health records with graph convolutional transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 606–613.
- Saibo Geng, Hudson Cooper, Michal Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. 2025. [JSONSchemabench: Evaluating constrained decoding with LLMs on efficiency, coverage and quality](#). In *ES-FoMo III: 3rd Workshop on Efficient Systems for Foundation Models*.
- 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.
- Diya Li, Yue Zhao, Zhifang Wang, Calvin Jung, and Zhe Zhang. 2024. [Large language model-driven structured output: A comprehensive benchmark and spatial data generation framework](#). *ISPRS International Journal of Geo-Information*, 13(11):405.
- Yaxi Lu, Haolun Li, Xin Cong, Zhong Zhang, Yesai Wu, Yankai Lin, Zhiyuan Liu, Fangming Liu, and Maosong Sun. 2025. [Learning to generate structured output with schema reinforcement learning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, pages 4905–4918, Vienna, Austria.
- OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Rimma Pivovarov and Noemie Elhadad. 2015. [Automated methods for the summarization of electronic health records](#). *Journal of the American Medical Informatics Association*, 22(5):938–947.
- Frank Schilder, Daniel Reichenpfader, et al. 2025. [Scientific evidence for clinical text summarization using large language models: Scoping review](#). *Journal of Medical Internet Research*, 27(1):e68998.
- Rosanne C Schoonbeek, Jessica D Workum, Stephanie C E Schuit, Anne H Hoekman, Taranom Mehri, Job N Doornberg, Tom P van der Laan, and Charlotte M H H T Bootsma-Robroeks. 2025. [Quality and efficiency of integrating customised large language model-generated summaries versus physician-written summaries: a validation study](#). *BMJ Open*, 15(9).
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atila 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, Lu Yang, Ke-jia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. 2025. [Medgemma technical report](#).
- Benjamin Shickel, Patrick J Tighe, Azra Bihorac, and Parisa Rashidi. 2018. [Deep ehr: A survey of recent advances in deep learning techniques for electronic health record \(ehr\) analysis](#). *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604.
- Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W Jim Jiménez, and Lucila Ohno-Machado. 2021. Deep representation learning of patient data from electronic health records (ehr): A systematic review. *Journal of Biomedical Informatics*, 115:103671.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, 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.
- Liyan Tang, Zhaoyi Sun, Betina Ilday, Jordan G Nestor, Ali Soroush, Noémie Elhadad, Chunhua Weng, and Yifan Peng. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.
- Dave Van Veen, Cara Van Uden, Louis Berry, and Yiliang and"; and"; and"; others Chen. 2024. [Adapted large language models can outperform medical experts in clinical text summarization](#). *Nature Medicine*, 30:1134–1142.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2018. [Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record](#). *Journal of the American Medical Informatics Association*, 25(3):230–238.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135.

Medical Text Rewriting for Non-Experts: A Guideline-Driven LLM Approach

Mana Kuramoto¹, Hiroyuki Nagai¹, Keiko Yamada², Hiroo Ide^{3,4}
Masayo Hayakawa^{3,5}, Tomohiro Nishiyama¹, Shoko Wakamiya¹, Eiji Aramaki¹

¹Nara Institute of Science and Technology, Japan

²Saitama Prefectural University, Japan ³The University of Tokyo, Japan

⁴Juntendo University, Japan ⁵Keio University, Japan

{kuramoto.mana.kj4, hiro.nagai}@naist.ac.jp, yamada-keiko@spu.ac.jp,

hiroo-ide@g.ecc.u-tokyo.ac.jp, hayakawam395@gmail.com,

{nishiyama.tomohiro.ns5, wakamiya, aramaki}@is.naist.jp

Abstract

Medical research is highly specialized, making it difficult for patients and general readers to understand recent findings. Traditionally, text simplification, replacing technical terms with more accessible expressions, has been employed. However, this approach alone is limited in addressing a lack of background knowledge and often results in the loss of important information. Therefore, this study defines “rewriting for non-experts” as a rewriting process that, in addition to simplification, supplements essential background knowledge such as the significance of the research and reasons it is needed and proposes a method for implementing this process using large language models (LLMs). To verify the effectiveness of the proposed approach, a quantitative evaluation using automatic metrics was conducted. The results showed that the method combining the guidelines for human text creation with few-shot examples of reference texts achieved the highest scores. The expansion of the guidelines is planned as part of future work to enable the rewriting of scientific and technological information in a form that is accessible to a broader audience.

Keywords: Medical Text Rewriting, Large Language Models, Comprehensibility Enhancement, Text Simplification, Lay Summarization, Non-expert Communication

1. Introduction

Scientific specialization has intensified the “silo effect,” isolating research fields and hindering cross-disciplinary communication (Tett, 2015). As expertise narrows and modern science becomes increasingly complex, even experts struggle to grasp developments outside their domains. Consequently, cutting-edge research remains largely inaccessible to laypeople. This accessibility gap is particularly critical in medicine: although open science grants patients access to primary literature, interpreting these texts without specialized training remains difficult (Gotlieb et al., 2022). Misinterpreting medical information poses significant risks, potentially disrupting patient–physician communication and compromising clinical decisions (Lu and Schulz, 2024; Root et al., 2016).

Current text simplification methods, which focus primarily on vocabulary substitution, fail to address deeper structural barriers (Alva-Manchego et al., 2021). The difficulty of scientific text stems not just from jargon, but from implicit assumptions, dense structure, and the reader’s lack of background knowledge. To bridge this gap, we introduce “rewriting for non-experts”—an advanced editing paradigm that, beyond simple simplification, involves selecting salient information, restructuring

the narrative, and clarifying the research’s significance. Fig. 1 illustrates this process: rather than merely simplifying vocabulary, the transformation explicates implicit context and foregrounds essential information while retaining necessary technical terms. This demonstrates that enhancing comprehensibility requires structural reorganization and contextual clarification.

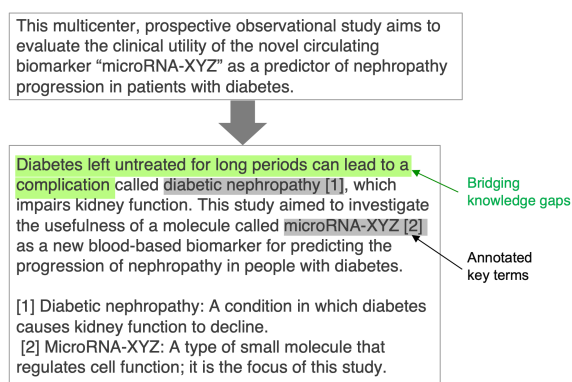


Figure 1: An example of rewriting for non-experts. The original expert-oriented text (top) is transformed into an accessible version (bottom) by explicating implicit context and restructuring the narrative.

As systematized in Fig. 2, our approach addresses three key barriers: (1) Implicit Knowledge: Unstated premises assumed by specialists. (2) Excessive Statistical Detail: Quantitative data that increases cognitive load without aiding non-expert understanding (Gigerenzer and Edwards, 2003). (3) Disrupted Flow: Frequent definitions that interrupt the narrative coherence (Leroy et al., 2013).

Consider a concrete scenario illustrating these barriers: a patient reads a press release about a promising new cancer drug. The text states that the drug “showed significant efficacy in clinical trials.” However, critical context—that the drug remains years from regulatory approval and is unavailable outside research settings—is omitted as self-evident to researchers. Such implicit assumptions can lead patients to request unavailable treatments, straining clinical relationships (Lu and Schulz, 2024; Root et al., 2016). This accessibility gap requires “rewriting for non-experts”—an advanced editing process that goes beyond simplification to include selecting key information, restructuring narratives, and explicating significance (Schmitz, 2023).

This study proposes a comprehensive rewriting framework that prioritizes comprehensibility while preserving essential medical accuracy. To this end, we investigate the feasibility of this complex task using Large Language Models (LLMs). Leveraging their capacity to integrate contextual information across discourse, the proposed approach employs LLMs to perform structural reorganization and context supplementation guided by human-authored references, distinguishing our approach from conventional simplification models.

2. Related Work

2.1. Lay Summarization

Lay summarization is the task of summarizing and restructuring documents containing specialized content into forms that are understandable to non-expert readers. For example, in the medical domain, a generative model that integrates background explanations and lexical substitutions has been proposed for producing summaries suitable for lay audiences (Guo et al., 2021). Furthermore, in the shared task BioLaySumm, a corpus aligning technical abstracts with lay summaries has been developed, along with a continuous evaluation framework (Goldsack et al., 2022, 2024).

Unlike conventional summarization, these studies are notable for aiming to generate documents tailored to the assumed knowledge levels and information needs of target readers. However, they face limitations, such as a narrow target audience and the absence of explicit modeling for appropri-

ately conveying the social significance and novelty of the research. For instance, in frameworks such as BioLaySumm and Explain-type LLMs, while readability and factual consistency are prioritized as evaluation criteria, the question of what kind of value is conveyed to whom is not explicitly addressed (Goldsack et al., 2024; Luo et al., 2024). Moreover, it has been reported that automatic evaluation metrics do not necessarily align with non-experts’ judgments of understandability and acceptability (Salvi et al., 2025), highlighting the need to redesign evaluation frameworks that consider reader attributes and perceptions.

2.2. Text Simplification

Text simplification is a task aimed at reducing readers’ cognitive load and enhancing comprehensibility by adjusting vocabulary and sentence structures in specialized or complex documents. In recent years, research has advanced across various domains, including news articles, academic writing, and medical texts. Automatic evaluation metrics play a central role in this field. SARI (Xu et al., 2016) is notable for its ability to independently assess additions, deletions, and retentions of lexical items. In contrast, metrics such as BLEU and FKGL (Kincaid et al., 1975) have been reported to exhibit important limitations for simplification evaluation, particularly in capturing simplicity and structural transformations (Alva-Manchego et al., 2021).

Applying these evaluation frameworks to the medical domain presents additional challenges. Medical documents rely heavily on technical terminology and domain-specific concepts, and even minor distortions may affect factual accuracy and safety. Consequently, simplification quality cannot be assessed solely through conventional readability-oriented metrics. Moreover, systematically aligned parallel corpora consisting of expert-level medical texts and their simplified counterparts remain limited. This scarcity complicates validation of automatic evaluation metrics in medical contexts, where meaning preservation and factual consistency are especially critical for rewriting for non-experts (Ondov et al., 2022).

Beyond domain-specific concerns, research has explored models that treat output difficulty as a controllable parameter (Martin et al., 2020), as well as controlled text simplification methods that explicitly specify target reader groups such as grade levels (Scarton and Specia, 2018). These approaches extend uniform simplification strategies by incorporating reader characteristics into the generation process. However, adjusting surface-level textual complexity does not necessarily resolve gaps in background knowledge between experts and non-experts. Empirical evaluations

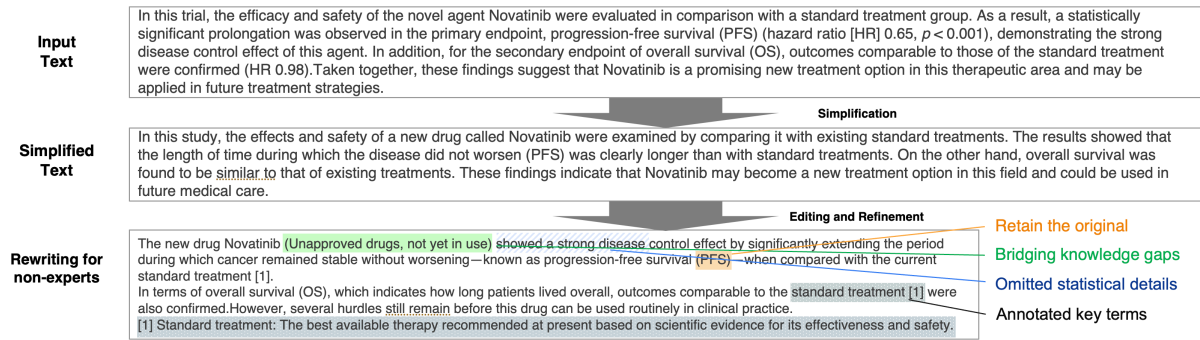


Figure 2: Overview of rewriting operations: supplementing implicit context (green), removing excessive details (blue), and explaining technical terms (orange).

based on reading comprehension tasks further indicate that important information may still be omitted during simplification (Agrawal and Carpuat, 2024). Taken together, these observations suggest that readability alone may not be sufficient to ensure adequate understanding of research background or significance.

2.3. Readability

Readability refers to quantitative measures of how easily a text can be processed by readers, and it has been widely used to evaluate the effectiveness of text simplification and lay summarization. Typical metrics rely on surface-level features such as lexical frequency, sentence length, and syntactic complexity (Kincaid et al., 1975). With the development of large-scale corpora for readability assessment, automatic metrics based on statistical modeling and machine learning have been proposed and applied to compare outputs of generative models (Crossley et al., 2022).

Although readability metrics provide useful indicators of linguistic difficulty, they primarily capture surface characteristics of texts. They do not directly account for how readers interpret information or construct meaning. In other words, even when lexical and syntactic complexity are reduced, abstract or conceptually dense content may remain difficult for readers without relevant background knowledge to understand. In rewriting for non-experts, therefore, formal readability constitutes a necessary condition, but it is not sufficient to ensure effective comprehension.

Recent studies have attempted to extend readability-oriented approaches by incorporating contextual generation processes. For example, multi-stage frameworks have been proposed in which large language models first generate background explanations and subsequently perform summarization (Luo et al., 2024). These approaches aim to enhance accessibility by supplementing contextual information while maintaining

linguistic simplicity.

Nevertheless, existing methods primarily focus on adding background information or adjusting textual difficulty at the surface level. They do not systematically reorganize information according to reader attributes or explicitly foreground research novelty and social significance. Thus, while readability remains a fundamental component, a more comprehensive framework that integrates document-level restructuring, contextual supplementation, and audience-sensitive organization remains underexplored. Within this landscape, rewriting for non-experts is positioned as a task that extends readability by emphasizing contextual reconstruction and explicit articulation of research significance for non-expert readers.

3. Dataset

This paper constructs a dataset consisting of pairs of simulated medical research summaries (hereafter, simulated summaries) and reference summaries manually rewritten by a medical writer. Because the copyright of the original reports belongs to individual researchers, we cannot directly release the original texts as a public dataset. To address this limitation and enable public release, we generated fictional summary documents (i.e., simulated summaries) using GPT-4.1-mini to mimic the content and writing style of the source reports.

To construct these simulated summaries, we used research outcome reports from AMEDfind¹, provided by the Japan Agency for Medical Research and Development (AMED), as source material. These simulated summaries focus on neoplasm-related clinical trials and drug development studies, which represent the most frequent category in the source data, and are used as the experimental source texts. The resulting pairs of simulated and reference summaries serve as the

¹<https://amedfind.amed.go.jp/amed/>

Item	Description
1. Overall policy	Target audience: Non-medical adult readers without prior knowledge of biology, equivalent to graduates of four-year undergraduate programs in the humanities and social sciences. Wordings should be selected so that readers without specialized knowledge can understand the research content and its key outcomes.
2. Structural and stylistic rules	The original writing style (plain or polite form) should be preserved. Reordering of words or phrases is allowed when necessary to improve clarity. Expressions or sentences that are unnecessary for understanding, such as sample size details or procedural descriptions, should be removed. When information is uncertain, definitive expressions should be avoided and appropriate hedging should be maintained.
3. Terminology rules	A plain expression should be presented first, followed by the original technical term in parentheses. Frequently used terms should be explained using footnotes. Abbreviations should be spelled out in full at their first occurrence. Chained possessives (e.g., "A of B of C") should be avoided by rephrasing or by adding explanatory footnotes.
4. Additional notes	Detailed examination procedures and analysis methods should be summarized concisely and described in plain language whenever possible.

Table 1: Guidelines for rewriting for non-experts

source and target texts for the two-stage rewriting method.

Manual rewriting for comprehensibility was then applied to the simulated summaries. Specifically, 20 documents were selected from the generated corpus, and one professional medical writer was commissioned to perform rewriting for non-experts. Since the simulated summaries were generated by LLMs, they contain no personally identifiable information or actual patient data. The resulting twenty rewritten documents were used as the reference summaries for evaluation.

4. Proposed Method

This paper proposes a two-stage LLM-based method for rewriting for non-experts. The method consists of sequential simplification and refinement stages.

In the Base stage, the LLM performs basic rewriting using prompts that instruct either simple simplification or direct rewriting for non-experts. In the Refinement stage, the output is further refined through prompts guided by predefined guidelines and few-shot examples, enabling more detailed restructuring and contextual adjustment. The two-stage configuration allows systematic comparison between surface-level simplification and guideline-informed refinement.

Base stage In the Base stage, two prompt types are employed to generate base texts for subsequent refinement.

Surface-level Simplification (Simp) : An instruction that focuses exclusively on reducing

linguistic complexity at the lexical and syntactic levels, without addressing contextual or background knowledge gaps. Example: *Please simplify the following text so that it can be understood by people without specialized knowledge.*

Comprehensive Rewriting (Direct) : An instruction that directly specifies rewriting for non-experts as the objective, which may involve not only simplification but also selection of key information, structural reorganization, and contextual clarification. The LLM is asked to produce a version appropriate for non-experts, allowing the model to determine the full range of strategies needed to achieve comprehensibility. Example: *Please perform rewriting for non-experts on the following text.*

Refinement stage The output generated in the Base stage is further refined by simulating the human rewriting process. In the proposed method, the prompt in this stage combines the following two components, both of which are applied together to guide the rewriting process.

Guidelines (GL) : A set of editing requirements independently designed and validated by experts. The detailed criteria are summarized in Table 2.3. These guidelines include rules for information supplementation and text structuring, such as adding explanations to technical terms and annotating frequently used terms outside the main text.

Few-shot examples (FS) : Example pairs consisting of original simulated summaries and

Input (Original)	非小細胞肺癌（NSCLC）において、EGFR 変異陽性例を対象に、既存治療薬に耐性を示す症例に対する新規の第三世代 EGFR チロシンキナーゼ阻害剤（TKI）の有効性および安全性評価を目的とした多施設共同第 II 相臨床試験を実施する。 A multicenter phase II clinical trial was conducted to evaluate the efficacy and safety of a novel third-generation EGFR tyrosine kinase inhibitor (TKI) in patients with non-small cell lung cancer (NSCLC) who were positive for EGFR mutations and had developed resistance to existing therapies.
Simp	非小細胞肺癌のうち、特定の遺伝子変異がある患者を対象に、これまでの薬が効かなくなった場合の新しい薬の効果と安全性を調べる臨床試験を、複数の施設で行いました。 Among patients with non-small cell lung cancer, a clinical trial was conducted at multiple institutions to examine the effectiveness and safety of a new drug in patients with specific gene mutations for whom previous treatments had become ineffective.
Simp+GLFS	非小細胞肺癌 [1] の患者さんの中には、特定の遺伝子 [2] に変化がある方がいます。これまでの薬が効かなくなってしまったそのような患者さんを対象に、新しい薬の効き目と安全性を調べる臨床試験を、複数の病院で行いました。 [1] 非小細胞肺癌：肺癌の一種で、比較的ゆっくりと進行するタイプです。 [2] 遺伝子：体の設計図のようなはたらきをする、細胞の中の物質です。 Among patients with non-small cell lung cancer [1], some have changes in specific genes [2]. In such patients for whom previous treatments had become ineffective, a multicenter clinical trial was conducted to evaluate the effectiveness and safety of a new drug. [1] Non-small cell lung cancer : A type of lung cancer that generally progresses relatively slowly. [2] Gene : A substance inside cells that functions like a blueprint for the body.
GLFS	肺癌の一種である非小細胞肺癌の患者のうち、特定の遺伝子（EGFR[1]）に変異がある人を対象に、臨床試験を行った。（中略）新しい第三世代の EGFR チロシンキナーゼ阻害剤（TKI）[2] という薬の効き目と安全性を評価することが目的だった。 [1] EGFR：細胞の増殖に関わるタンパク質（受容体）の一種。この遺伝子に変異があると、がん細胞が増えやすくなる。 [2] EGFR チロシンキナーゼ阻害剤（TKI）：EGFR というタンパク質のはたらきを抑えることで、がん細胞の増殖を止める薬。 A clinical trial was conducted in patients with non-small cell lung cancer, a type of lung cancer, who had mutations in a specific gene, EGFR [1]. ... The objective was to evaluate the efficacy and safety of a new third-generation EGFR tyrosine kinase inhibitor (TKI) [2]. [1] EGFR: A type of receptor protein involved in cell growth; mutations in this gene can promote the proliferation of cancer cells. [2] EGFR tyrosine kinase inhibitor (TKI): A drug that suppresses cancer cell growth by inhibiting the activity of the EGFR protein.

Table 2: Comparison of generation examples produced by DeepSeek-V3.2. Note that English translations are provided for the generated Japanese text. Under the two-stage setting, important technical terms such as EGFR are omitted, whereas the single-stage setting preserves these terms while providing additional explanations.

their corresponding reference summaries are provided to illustrate how the guidelines are applied in practice. Through in-context learning (ICL), the LLM acquires concrete patterns for generating text consistent with rewriting for non-experts.

5. Experiments

To evaluate the effectiveness of the proposed two-stage rewriting method, comparative experiments using LLMs were conducted. Simulated summary documents were used as input, and the rewritten texts were quantitatively evaluated based on their similarity to the reference summaries produced by

a medical writer.

The experiments analyze the impact of two variables on output quality: (1) the type of instruction used in the Base stage, namely simplification or direct rewriting for non-experts, and (2) the presence or absence of refinement using additional information, namely guidelines and few-shot examples, in the Refinement stage.

All texts generated under each experimental setting are publicly available.²

²<https://github.com/sociocom/E2U>

Model	Setting		Similarity Metrics		Summary Metrics		
	Base	Refinement	BERTScore	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
DeepSeek	Simp	–	0.774	16.827	0.543	0.244	0.401
	Direct	–	0.766	15.394	0.563	0.240	0.382
	Original	GL	0.747	11.145	0.742	0.327	0.311
	Original	GLFS	0.773	19.607	0.763	0.403	0.398
	Direct	GL	0.745	12.764	0.641	0.256	0.339
	Direct	GLFS	0.760	15.148	0.627	0.270	0.360
	Simp	GL	0.748	12.743	0.647	0.266	0.337
	Simp	GLFS	0.762	16.917	0.667	0.308	0.375
Gemini	Simp	–	0.771	17.540	0.633	0.282	0.393
	Direct	–	0.762	15.407	0.648	0.276	0.376
	Original	GL	0.730	8.894	0.777	0.361	0.251
	Original	GLFS	0.753	13.499	0.476	0.248	0.318
	Direct	GL	0.729	9.161	0.727	0.303	0.280
	Direct	GLFS	0.748	12.600	0.722	0.315	0.333
	Simp	GL	0.755	11.079	0.728	0.325	0.296
	Simp	GLFS	0.749	13.023	0.738	0.346	0.321
GPT-5.2	Simp	–	0.784	19.288	0.591	0.276	0.420
	Direct	–	0.777	18.596	0.617	0.286	0.402
	Original	GL	0.750	12.279	0.750	0.347	0.309
	Original	GLFS	0.753	13.060	0.796	0.415	0.307
	Direct	GL	0.747	13.655	0.705	0.313	0.329
	Direct	GLFS	0.754	15.040	0.704	0.324	0.356
	Simp	GL	0.753	13.123	0.711	0.320	0.325
	Simp	GLFS	0.755	14.322	0.728	0.338	0.343

Table 3: Automatic evaluation results under various rewriting settings. For each model, the best value for each evaluation metric is shown in bold. The shaded rows indicate the proposed two-stage method (Simp+GLFS).

5.1. Models

The following three LLMs were selected for the experiments based on their demonstrated effectiveness in prior text generation and simplification studies:

- **DeepSeek-V3.2** (DeepSeek-AI, 2025): A reasoning-focused language model developed by DeepSeek. Released with open weights and known for cost efficiency.
- **Gemini 2.5 Flash** (Gemini Team, 2025): A multimodal language model developed by Google. Optimized for efficiency and response speed.
- **GPT-5.2 (gpt-5.2-2025-12-11)** (OpenAI, 2025): A general-purpose language model developed by OpenAI. Widely used in natural language processing research.

For parameter settings, the temperature was set to 0.0 and the maximum output length to 2048 tokens for DeepSeek-V3.2 to ensure stable generation. For Gemini 2.5 Flash and GPT-5.2, the default parameter settings provided by each API were used without further modification.

5.2. Evaluation Data and Metrics

For evaluation, 20 pairs of simulated summary documents and their corresponding reference summaries were used, as described in Section 3. In this experiment, two randomly selected pairs were used as few-shot examples in the Refinement stage, and the remaining 18 pairs were used as test data.

To assess the quality of the generated texts from multiple perspectives, **BERTScore** was used to measure semantic similarity, **BLEU** to assess surface-level overlap, and **ROUGE-1/2/L** to evaluate n-gram and sequence-level agreement.

5.3. Ablation Settings

To examine the contribution of each component of the proposed method to rewriting performance, ablation experiments were conducted by varying the combination of instructions in the Base stage (Simp/Direct) and the additional information provided in the Refinement stage (GL/GLFS).

Input	A newly identified transcription factor, XZF-1, has been shown by recent basic research to play an important role in the development and progression of lung cancer. (omitted) Furthermore, based on the design of early-phase clinical trials to evaluate the efficacy of XZF-1 inhibitors, protocols are being developed to establish a foundation for their future introduction as standard treatment. 新規に同定された転写因子 XZF-1 は、肺がんの発生および進展に重要な役割を果たすことが近年の基礎研究で示された。(中略)さらに、XZF-1 阻害剤の有効性を評価する初期臨床試験設計に基づき、将来的な標準治療への導入に向けたプロトコル策定を進めている。
GPT-5.2	A regulatory factor called “XZF-1,” which controls gene activity, has recently been shown through experimental research to be involved in the processes by which lung cancer develops and progresses. (omitted) Based on these findings, trial protocols are being prepared to evaluate the effects of XZF-1 inhibitors, with the aim of establishing a foundation for their future inclusion as a standard treatment option. Number of annotations: 7 (limited to key technical terms) 新しく見つかった「XZF-1」という遺伝子のはたらきを調整する因子(転写因子)は、肺がんができたり進んだりする過程に関わっている可能性が、近年の実験中心の研究で示されてきた。(中略)こうした結果をもとに、XZF-1 阻害剤の効果性を評価する試験計画を整え、将来的に標準的な治療選択肢として導入するための基盤づくりを進めている。
Gemini	Recent basic research has shown that XZF-1, a newly discovered transcription factor, plays an important role in how lung cancer begins and progresses. (omitted) In addition, preparations are underway to create a detailed protocol for early-stage clinical trials to evaluate the efficacy of XZF-1 inhibitors, with the goal of introducing them in the future as a treatment that is currently established as the most effective and safe (standard treatment). Number of annotations: 22 (including common terms such as “lung cancer” and “efficacy”) 新たに発見された転写因子 XZF-1 が、肺がんの発生や進行において重要な役割を果たすことが基礎研究により示された。(中略)さらに、XZF-1 阻害剤の有効性を評価する初期臨床試験の詳細な計画(プロトコル)を策定し、将来的に標準治療として導入するための準備を進めている。

Table 4: Comparison of generation outputs between GPT-5.2 and Gemini under the Original+GLFS setting. GPT-5.2 produces 7 annotations limited to key technical terms, whereas Gemini generates approximately three times more annotations (22), including common terms such as “lung cancer” and “efficacy.”

6. Results and Discussion

Table 3 presents the automatic evaluation results. Prompt design had a notable effect on the scores. The single-stage setting (**Original+GLFS**) outperformed the two-stage setting (**Simp+GLFS**) across most metrics. GPT-5.2 achieved the highest BLEU scores, while DeepSeek and Gemini showed greater variation across settings.

6.1. Single-stage vs. Two-stage Approaches

The two-stage approach yielded lower performance. The Base stage simplifies technical terms; for instance, “EGFR mutation” becomes “specific gene mutations” (Table 3). Once simplified, these terms are difficult to recover in the Refinement stage, leading to information loss.

In contrast, the single-stage approach retains original terminology while adding explanations simultaneously. This preserves both accuracy and readability. These results indicate that direct generation is more effective than stepwise processing for specialized content.

6.2. Model Comparison

GPT-5.2 achieved the highest BLEU scores by making minimal structural changes and limiting annotations to essential terms, consistent with guideline item 4. Gemini, however, produced approximately three times more annotations, including explanations for common terms such as “lung cancer” (Table 3). This verbosity reduced BLEU scores due to lower precision, though ROUGE scores remained stable as the core content was preserved. Controlling annotation density according to reader characteristics remains a key challenge.

6.3. Effect of Few-shot Examples

Comparing Original+GL with Original+GLFS reveals the contribution of few-shot examples to rewriting quality. BLEU scores improved substantially when few-shot examples were added: DeepSeek showed a 75% increase (11.1 to 19.6), Gemini 52% (8.9 to 13.5), and GPT-5.2 6% (12.3 to 13.1). BERTScore also improved consistently across all models, indicating enhanced semantic alignment with reference texts.

These results suggest that guidelines alone may

be insufficient for LLMs to infer the expected output format and annotation style. Few-shot examples provide concrete demonstrations of how guidelines should be applied, reducing ambiguity in interpretation. The improvement was most pronounced for DeepSeek, suggesting that models with stronger in-context learning capabilities may benefit more from example-based guidance.

Interestingly, Gemini exhibited a decrease in ROUGE scores under GLFS compared to GL alone (ROUGE-1: 0.78 to 0.48), despite improvements in BLEU and BERTScore. This discrepancy may reflect Gemini’s tendency to generate additional explanatory content when provided with examples, which increases output length and reduces n-gram overlap with the reference.

These findings also point to directions for future research. Constructing a richer reference dataset authored by multiple writers would enable a more objective evaluation that does not rely on the writing style or judgment of a single individual. Furthermore, human evaluations involving actual patients and general readers are essential to assess guideline adherence and confirm the absence of medical misunderstandings.

7. Conclusion

Rewriting for non-experts was defined herein as a framework for presenting medical research texts in a form that is accessible to non-expert readers, and its feasibility was investigated using LLMs. The results suggest that directly generating the final output through editing guidelines and concrete examples is effective for achieving rewriting for non-experts.

The goal of the concept targeted in this work is not only to facilitate communication among experts but also to support the creation of a society in which people from diverse backgrounds can access and understand scientific and technological information, including medical knowledge, in a clear and inclusive manner. Future work includes constructing larger datasets covering a broader range of disease domains and exploring mechanisms to ensure the accuracy of medical information.

8. Limitations

This study has several limitations that should be acknowledged. First, the reference texts used for evaluation were created by a single medical writer. There is no single correct answer for rewriting for non-experts; the optimal level of detail and phrasing may vary depending on the assumed knowledge level of the reader and the reading context.

Therefore, the current evaluation, based on the degree of alignment with a single writer’s output, captures only one dimension of performance.

Second, limitations are inherent in the automatic evaluation metrics used. This study employed n-gram-based metrics such as BLEU and ROUGE, which assess surface-level similarity to reference texts but do not directly measure understandability or adherence to guidelines. For example, in our experiments, Gemini received lower scores due to its tendency to add excessive annotations, which deviated from the guidelines. This outcome is consistent with the behavior of these metrics in penalizing verbose or off-target output. However, even if such annotations were beneficial to readers, they would still be penalized unless included in the reference. Thus, current metrics make it difficult to distinguish between unnecessary redundancy and useful supplementary information.

Regarding the reviewer’s question of whether non-expert human evaluation is necessary, we consider it essential for this task. Because rewriting for non-experts targets perceived clarity, trust, and actionable understanding, automatic overlap metrics alone are insufficient to establish practical usefulness. In this study, we prioritized controlled comparisons across prompting settings as an initial step; however, a follow-up evaluation with non-expert participants is required to assess comprehensibility, perceived safety, and guideline appropriateness. We plan to include such human evaluation as a core component of the next experimental phase.

9. Acknowledgments

This study was supported by AMED Grant Number JP25oa0439009, the Strategic Innovation Promotion Program (SIP) “Development of an Integrated Healthcare System” JPJ012425, and JSPS Research Start-up Support JP25K24412.

10. Bibliographical References

- Sweta Agrawal and Marine Carpuat. 2024. *Transactions of the Association for Computational Linguistics*, 12:432–448.
- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. [The \(un\)suitability of automatic evaluation metrics for text simplification](#). *Computational Linguistics*, 47(4):861–889.
- Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2022. [A large-scaled corpus for](#)

- assessing text readability. *Behavior Research Methods*, 55:1023–1047.
- DeepSeek-AI. 2025. *Deepseek-v3.2: Pushing the frontier of open large language models*. *arXiv preprint arXiv:2512.02556*.
- Gemini Team. 2025. *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*. Technical report, Google. Accessed: 2026-01-08.
- Gerd Gigerenzer and Adrian Edwards. 2003. *Simple tools for understanding risks: From innumeracy to insight*. *BMJ: British Medical Journal*, 327(7417):741–744.
- Thomas Goldsack, Seán Cartright, Jon Chamberlain, and Massimo Poesio. 2022. *Making science simple: Corpora for the lay summarisation of scientific literature*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Thomas Goldsack, Jon Chamberlain, and Massimo Poesio. 2024. *Overview of the BioLay-Summ 2024 shared task on the lay summarization of biomedical research articles*. In *Proceedings of the 2024 BioNLP Workshop*, Bangkok, Thailand. Association for Computational Linguistics.
- Rachael Gotlieb, Corinne Praska, Marissa A. Hendrickson, Jordan Marmet, Victoria Charpentier, Emily Hause, Katherine A. Allen, Scott Lunos, and Michael B. Pitt. 2022. *Accuracy in patient understanding of common medical phrases*. *JAMA Network Open*.
- Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. *Automated lay language summarization of biomedical scientific reviews*. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):160–168.
- 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*. Research Branch Report 8-75, Chief of Naval Technical Training, Naval Air Station Memphis, Memphis, TN.
- Gondy Leroy, David Kauchak, and Obay Mouradi. 2013. *A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty*. *International Journal of Medical Informatics*, 82(8):717–730.
- Qing Lu and Peter J. Schulz. 2024. *Physician perspectives on internet-informed patients: Systematic review*. *Journal of Medical Internet Research*, 26(1):e47620.
- Ziqi Luo, Shervin Radpour, Shihua Wang, Sarah Yeston, Yining Ruan, and Mark Gerstein. 2024. *The lay person's guide to biomedicine: Orchestrating large language models*. *arXiv preprint arXiv:2402.13498*.
- Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. *Controllable sentence simplification*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.
- Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. *A survey of automated methods for biomedical text simplification*. *Journal of the American Medical Informatics Association: JAMIA*, 29(11):1976–1988.
- OpenAI. 2025. *Introducing gpt-5.2*. Accessed: 2026-01-08.
- Jessica Root, Nancy V. Oster, Shanda L. Jackson, Ralitza Mejilla, Jan Walker, and Joann G. Elmore. 2016. *Characteristics of patients who report confusion after reading their primary care clinic notes online*. *Health Communication*, 31(6):778–781.
- Rahul C. Salvi, Tanvir Alam, Sambuddha Ghosh, Hrishikesh Hegde, Soham Ghosh, Alexis Palmer, and Danielle Mowery. 2025. *Towards understanding LLM-generated biomedical lay summaries*. In *Proceedings of the 2025 Workshop on Clinical Natural Language Processing for Health (CL4Health)*, Mexico City, Mexico. Association for Computational Linguistics.
- Carolina Scarton and Lucia Specia. 2018. *Learning simplifications for specific target audiences*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.
- Birte Schmitz. 2023. *Improving accessibility of scientific research by artificial intelligence—an example for lay abstract generation*. *Digital Health*, 9:20552076231186245.
- Gillian Tett. 2015. *The Silo Effect: The Peril of Expertise and the Promise of Breaking Down Barriers*. Simon & Schuster, New York.

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.

Patient-Specific Care Pathway Visualisation for Medical Nursing Staff

Sophie Bauernfeind, Selina Adlberger, Oliver Krauss, Andreas Stöckl

Research & Development, Research & Development, Digital Media Lab, Digital Media Lab,
University of Applied Sciences Upper Austria

Hagenberg, Austria

{sophie.bauernfeind, selina.adlberger, oliver.krauss, andreas.stoeckl}@fh-hagenberg.at

Abstract

Nursing staff are increasingly confronted with extensive and detailed patient documentation, requiring much time to read through numerous possible care measures. Combined with rising patient loads, this underscores the need for a clearer and more immediately accessible overview of each patient's situation. Patient-specific care pathway visualisations offer a promising approach to reduce cognitive load, support faster decision-making, and improve situational awareness. This work investigates two Artificial intelligence (AI)-assisted methods for generating such visualisations: (1) simple image generation based on structured textual prompts, and (2) automated code generation that produces graph-based representations of clinical pathways. Using a dataset of synthetic patient profiles and seven defined care pathways, evaluating multiple state-of-the-art foundation models. The results highlight clear differences between models and approaches, particularly in language sensitivity, structural consistency, and the level of detail achievable. Image-based outputs provided visually rich overviews but frequently introduced subtle logical inconsistencies, while code-based methods produced verifiable and structurally coherent pathways yet varied in their ability to preserve contextual and psychosocial information. Together, these findings indicate that AI-assisted visualisation can effectively support—but not yet fully automate—patient-specific pathway generation, and they point toward hybrid solutions that combine visual accessibility with logical robustness.

Keywords: care pathways, clinical decision support, generative AI, visualization, elderly care

1. Introduction

In modern healthcare systems, nursing staff face increasing pressure as they manage growing patient loads, complex care requirements, and rapidly changing clinical environments (Jennings et al., 2022; Rachel and Francesco, 2018). A substantial portion of the nursing workload arises from the considerable time required to review extensive patient documentation and identify the most appropriate next steps in care. Nurses are frequently required to extract and synthesize information from numerous, often lengthy records to construct an accurate and up-to-date understanding of a patient's condition.

Visualisation and visualizing patient-specific care pathways has emerged as a promising strategy to address these challenges (Stadler et al., 2016; Mould et al., 2010; Tan et al., 2025). Instead of relying on static or generic process models, patient-specific visualizations dynamically reflect the individual's medical history, current interventions, and expected future steps. Creating these tailored visualizations remains a non-trivial task. Clinical pathways are complex, vary between patients, and evolve over time, which makes automated, flexible, and context-sensitive generation essential.

In this work, methods for generating patient-specific care pathway visualizations that can be seamlessly integrated into nursing

workflows are investigated. Specifically, two methodological approaches got explored: (1) simple image generation, which allows creation of illustrations without requiring technical expertise, and (2) automated code generation, which produces graph representations using visualization libraries. The former prioritizes accessibility, enabling staff to obtain illustrative overviews with minimal effort, while the latter provides precision, adaptability, and integration potential with clinical data systems.

By comparing these two approaches, a better understanding is sought of how Artificial intelligence (AI)-assisted visualization generation can support nursing staff in managing complex care processes.

2. Related Work

Patient journey mapping provides structured visual descriptions of patients' movements through the healthcare system, helping clinicians identify barriers, transitions, and needs across the care continuum (Bulto et al., 2024). Such representations offer an important foundation for developing individualized visualizations that support rapid comprehension (Wagenaar et al., 2026).

Studies on pathway modeling emphasize the role of diagrammatic representations in clinical process improvement. Mould et al. demonstrate that accessible pathway diagrams can assist staff in identify-

ing inefficiencies and evaluating redesign options, thereby improving communication and supporting organizational learning (Mould et al., 2010). Their findings highlight the practical relevance of clear visual models in everyday clinical workflows.

Gartner et al. propose an integrative framework that accounts for clinical, organizational, and informational dimensions, underscoring the need for pathway representations that reflect the complex realities of patient care and interprofessional collaboration (Gartner et al., 2022).

Manktelow et al. summarize methods for extracting time-resolved sequences of care activities from administrative and clinical data, showing how empirical pathway structures can support planning, quality improvement, and decision-making (Manktelow et al., 2022). These approaches demonstrate the potential for automated generation of patient-specific pathways grounded in actual care delivery.

Existing work on generative graph and visualization models has demonstrated strong performance, but these approaches typically rely on datasets that are already visualized or highly pre-structured. Automated data-to-visualization systems, for example, operate on clean tabular inputs that can be directly mapped into plotting code rather than inferred from narrative descriptions (Wu et al., 2024; Zou et al., 2024; Wang et al., 2025).

3. Methods

This study investigates two complementary approaches for generating patient-specific care pathway visualizations that support nursing staff in rapid situation assessment: (A) simple image generation and (B) automated code generation for graph-based visualizations. Both approaches are evaluated with respect to feasibility, interpretability, adaptability, and potential for integration into clinical workflows. To ensure a fair comparison, a shared data foundation and a common set of evaluation criteria are employed.

3.1. Care Pathway

A total of seven distinct clinical guidelines are utilized, each comprising between 5,771 and 126,403 words. These pathways encompass the following thematic areas: (1) Work Aid, (2) decubitus prophylaxis, (3) Dementia, (4) Urinary Incontinence, (5) Chronic Wound Care, (6) Pain Management, and (7) Fall Prevention.

3.2. Patient Data

In this study, a dataset consisting of ten synthetic patient profiles were generated with GPT-4. Each profile is provided as a plain text document structured around fourteen information categories.

These categories encompass: (0) general patient information such as name, gender, and age; (1) communication abilities, including hearing, vision, speech, and social interaction; (2) mobility, covering aspects such as movement in bed, and use of assistive devices; (3) vital functions including blood sugar management, medication handling, and pain assessment; (4) self-care activities, ranging from independent or assisted hygiene routines to skin and oral care; (5) eating and drinking behaviors, including levels of independence and support required for meal preparation and intake; (6) toileting abilities, continence status, and the use of catheters or related aids; (7) dressing and undressing capabilities, including the need for support and use of prostheses or orthoses; (8) sleep patterns, disturbances, and related observations; (9) hobbies and leisure activities; (10) preferences regarding the gender of caregiving staff as well as individual care goals and planned interventions; (11) safety-related aspects, such as orientation and the ability to recognize hazards; (12) household-related competencies, including heating, pet care, cleaning responsibilities, and relevant observations; and (13) personal expectations and wishes, religious considerations and preferences of both patients and their trusted persons, shown in Listing 1.

This dataset provides a consistent basis for comparing visualization approaches across varied patient situations, with documents ranging from 1,631 to 1,917 words.

3.3. Data Preparation and Pre-processing

Given the variability in patient data, not all treatment guidelines are applicable to every individual case. To streamline processing and ensure the accurate generation of visualization graphs, a two-stage pre-processing pipeline to structure both the clinical logic and the patient-specific relevance was implemented.

Guideline Structuring (Rule Extraction). Original clinical guidelines are typically available as narrative text, which can lead to ambiguity during automated processing. To mitigate this, Gemini 3 Pro was used to transform the raw guideline texts into a JavaScript Object Notation (JSON) format. Since the source guidelines were provided in German, the extracted logic retains the original German medical terminology to prevent semantic loss during translation. In this file, complex care instructions were decomposed into atomic "IF-THEN" statements:

- **Condition (IF):** Trigger criteria based on patient assessment (e.g., "IF Braden score is low").

First name: Elisabeth
 Last name: Müller
 Address: Musterstraße 10, 1010 Vienna
 Date of birth: April 12, 1935
 Main diagnoses: Dementia, high blood pressure, osteoarthritis
 [...]
 Care level: 4, application for increase submitted: no
 Emergency call system: yes, with Johanniter Unfallhilfe
 [...]
 Speech disorder: No
 Type of speech disorder: -
 Skills/resources/
 Observations: Despite her dementia, Elisabeth is still able to hold simple
 → conversations. She responds positively to friendly and patient communication.
 → She tends to repeat herself and has difficulty remembering new situations and
 → faces. She shows confusion when discussing complex topics. She likes being read
 → to and finds it relaxing.

2. ABEDL Moving around Independently Supervision Instruction Partial
 → assistance Full assistance Relevant for care plan.

Movement in bed (changing position/sitting up): X
 [...]
 Bowel movements: regular irregular X normal diarrhea constipation
 [...]

Listing 1: Excerpt of a patient profile, translated to provide reader clarity. Full original record contains 1,773 words.

- **Action (THEN):** The mandatory nursing intervention (e.g., "THEN initiate positioning schedule").

This transformation reduces the input tokens for the language model by converting text into a clearer and cleaner structure. An excerpt of the generated structure is shown in [Listing 2](#).

Model	A1	A2	B1	B2
Gemini 3 Pro	X	X	X	X
Gemini 3 Fast	-	-	X	X
Gemini 3 Thinking	X	X	-	-
NanoBanana Pro	X	X	-	-
GPT-5	-	-	X	-

Table 1: Model usage across test setup.

Patient-Guideline Mapping. Since applying all guidelines to every patient would create cluttered and irrelevant visualizations, an intermediate mapping step was introduced. Gemini 3 Pro was used to automatically generate a specific mapping file that acts as a relevance filter. The model analyzed the unstructured patient profile to identify relevant diagnoses and care needs, subsequently flagging specific guideline modules as `active: true` or `active: false` (e.g., activating "Dementia Care" only for patients with a confirmed diagnosis). This ensures that the generated pathways focus exclusively on clinically indicated interventions.

3.4. Experimental Models

To evaluate the two visualization approaches, state-of-the-art foundation models were employed. Models were allocated to Approach A (Image Generation) and Approach B (Code Generation) as shown in [Table 1](#). All models were tested with prompts in both English and German to evaluate cross-lingual robustness.

Experimental Design Framework. To systematically assess both visualization approaches, an evaluation framework was developed. Within this framework, three key parameters were varied:

- **Data Source:** Pre-processed Logic (JSON) vs. Unstructured Full-Text (All Guidelines).
- **Model:** Depending on the approach, models are used as listed above.
- **Prompt Language:** German (Native Source) vs. English (Translation).

3.5. Approach A: Simple Image Generation

The first approach explores the capability of multi-modal models to synthesize complex clinical data directly into visual artifacts, bypassing traditional rendering software. To ensure clinical utility, strict, role-based prompts were designed to enforce specific aesthetic and structural constraints (see [Listing 3](#) and [Listing 4](#)).

```

{
"arbeitshilfe_complete_txt": [
  "IF Aufnahme eines neuen Patienten THEN Durchführung der SIS (Strukturierte
  ↳ Informationssammlung) innerhalb von 24h.",
  "IF Risikoeinschätzung (Matrix) zeigt Handlungsbedarf THEN Planung konkreter
  ↳ Manahmen im Manahmenplan.",
  "IF Manahmen durchgeführt THEN Dokumentation im Berichtsblatt (Abweichungen
  ↳ begründen).",
  "IF Evaluationsdatum erreicht OR akute Zustandsnderung THEN Erneute Einschätzung
  ↳ der Risiken und Anpassung des Manahmenplans.",
  "IF Pflegegrad 4 oder 5 THEN Tgliche Dokumentation der Krperpflege und Mobilität
  ↳ zwingend erforderlich."
]
}

```

Listing 2: Excerpt of the pre-processed guideline logic in german. Explicit IF-THEN rules were extracted by Gemini 3 Pro to standardize the input for downstream visualization.

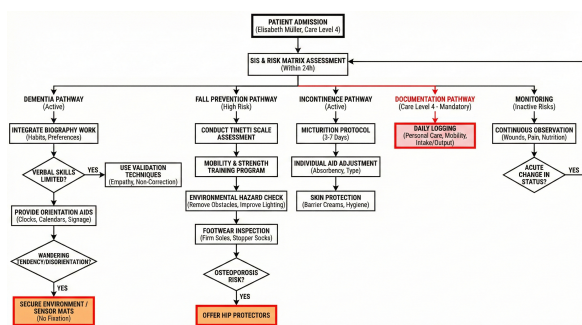


Figure 1: Generated image from NanoBanana Pro, with English prompt.

3.5.1. Input Variation A1: Structured Guideline Injection

In this setup, the model used pre-processed “If-Then” rules (`guidelines.json`) and a patient filter (`guidelines_patients.json`) to generate a decision tree (see Listing 3). The objective was to test whether providing structured logic improves the accuracy and consistency of the visual pathways.

3.5.2. Input Variation A2: Unstructured Full-Text Injection

In this variation, the model received the complete, unprocessed textual guidelines without specific visual constraints (see Listing 4). The objective was to assess its ability to autonomously extract clinical rules from raw text, resolve ambiguities, and generate coherent pathways.

Image Synthesis. The visualization process was orchestrated by Gemini 3 Pro, utilizing the NanoBanana Pro model as the generative engine within the Gemini environment (see Figure 1). Gemini 3 Pro first processed the prompt and patient data to structure the visual request, subsequently instructing NanoBanana Pro to generate

System Role: You are a technical
 ↳ illustrator for medical algorithms.
 ↳ Create a highly detailed
 ↳ visualization of a clinical pathway
 ↳ as a flowchart.

HERE IS THE DATA:

1. PATIENT (The focus):[Insert
 ↳ contents of 'patient_0.txt']
 ### 2. ACTIVE TOPICS (The
 ↳ filter):[Insert contents of
 ↳ 'guidelines_patients.json']
 ### 3. RULES (The logic):[Insert
 ↳ contents of 'guidelines.json']

IMAGE INSTRUCTIONS:

- Generate a strict "flowchart" on a
 ↳ white, technical background.
- Show clearly defined nodes (boxes) and
 ↳ connecting lines (arrows).
- Structure: At the top is the starting
 ↳ point "patient admission," below
 ↳ which the tree branches into
 ↳ different paths (fall, dementia,
 ↳ care).
- Style: Vector graphics, schematic,
 ↳ clean, medically precise. No photos,
 ↳ no 3D effects.
- Colors: Black/white with signal colors
 ↳ (red/orange) for risk paths.

Additional note for AI: Try to suggest
 ↳ text in the boxes, but focus
 ↳ primarily on the correct logical
 ↳ branching aesthetics of a decision
 ↳ tree.

Listing 3: Prompt template for Approach A1. The model receives structured JSON rules and specific color signaling instructions to generate the visualization.

the output. The model was instructed to emphasize a decision tree’s logical branching, placing text in boxes when possible while preserving the correct care-path topology.

System Role: You are a technical
 ↳ illustrator specializing in medical
 ↳ algorithms. Create a detailed
 ↳ visualization of a clinical pathway
 ↳ in the form of a flowchart.
 HERE IS THE DATA:
 ### 1. PATIENT (The Focus):[Insert
 ↳ contents of 'patient_0.txt']
 ### 2. ACTIVE ISSUES:[Insert contents of
 ↳ 'guidelines_patients.txt']
 ### 3. RULES (The Logic):[Insert
 ↳ unstructured raw guideline texts]
 IMAGE INSTRUCTIONS:
 - Create a strict "flowchart" on a
 ↳ white, technical background.
 - Show clearly defined nodes (boxes) and
 ↳ connecting lines (arrows).
 - Structure: At the top is the starting
 ↳ point "Patient Admission"; below
 ↳ that, the tree branches into
 ↳ different paths (Fall, Dementia,
 ↳ Care).
 - Style: Vector graphics, schematic,
 ↳ clear, medically precise. No photos,
 ↳ no 3D effects.

Listing 4: Prompt template for Approach A2. The model must extract visual logic directly from unstructured text guidelines without specific color signaling constraints.

Quality Assurance Criteria. Generated outputs were screened against:

1. **Risk Visibility:** Whether patient-specific risk factors (e.g., fall risk, dementia-related supervision needs) were clearly represented and correctly highlighted according to the defined visual signaling scheme.
2. **Data Faithfulness:** Whether the visual branching logic corresponded to the injected patient data and guideline content, without omission of central care elements.
3. **Occurrence of Hallucinated Elements:** Whether the visualization contained invented nodes, incorrect causal links, dangling edges, or visual artifacts not grounded in the provided input data.

3.6. Approach B: Code-Generated Graph Visualizations

The second approach prioritizes interoperability and lightweight system integration by generating Mermaid.js code (Sveidqvist and the Mermaid.js Contributors, 2014). Unlike complex rendering scripts, Mermaid offers a declarative, markdown-compatible syntax that allows care pathways to be embedded directly into clinical dashboards as renderable text.

3.6.1. Input Variation B1: Structured Logic

In this setup, Gemini 3 Pro, Gemini 3 Fast, and GPT-5 received the guidelines.json containing explicit "If-Then" rules alongside the patient profile (see Listing 5). The objective was to test the models' ability to strictly follow syntax constraints and generate valid Mermaid flowchart syntax when the logic extraction is already pre-processed.

System Role: You are a medical
 ↳ assistant. Create a precise process
 ↳ graph (Mermaid.js) for care
 ↳ planning.
 Use these three data sources:
 ### SOURCE 1: THE PATIENT (status &
 ↳ vitals): [Insert contents of
 ↳ 'patient_0__full.txt']
 ### SOURCE 2: RELEVANT TOPICS (Filter):
 ↳ [Insert contents of
 ↳ 'guidelines_patients.json']
 ### SOURCE 3: LOGIC & RULES (Knowledge
 ↳ Base): [Insert contents of
 ↳ 'guidelines.json']
 YOUR TASK: Create a flowchart that
 ↳ logically deduces what needs to be
 ↳ done.
 Procedure:
 1. Look in SOURCE 2: Which topics are
 ↳ active? (e.g., falls, dementia).
 2. Go to SOURCE 3 (rules) for these
 ↳ topics.
 3. Use the data from SOURCE 1 (patient)
 ↳ to check which "if" condition is
 ↳ met.
 - Example: Rule says "If Braden <
 ↳ 12." According to Source 1, the
 ↳ patient has "Braden 10." ->
 ↳ Include the measure "soft
 ↳ bedding" in the graph.
 - If a condition is NOT met, omit the
 ↳ measure.
 Just give me back the Mermaid code.

Listing 5: Prompt template for Approach B1. The model receives pre-extracted JSON rules and a concrete deduction example.

3.6.2. Input Variation B2: Unstructured Full-Text

In this setup, only the Gemini 3 Series models were tested. GPT-5 was excluded from this specific variation due to platform-imposed file upload limits that prevented the ingestion of the complete guideline set. As shown in Listing 6, the models had to perform implicit logic extraction directly from raw text. The objective was to evaluate the reasoning capabilities of the *Pro* versus *Fast* variants when filtering noisy, real-world data and structuring it into a valid Mermaid graph within a single context window.

```

System Role: You are a medical
→ assistant. Create a precise process
→ graph (Mermaid.js) for care
→ planning.
Use these data sources:
### SOURCE 1: THE PATIENT (status &
→ vitals): [Insert contents of
→ 'patient_0__full.txt']
### SOURCE 2: RELEVANT TOPICS (Filter):
→ [Insert contents of
→ 'guidelines_patients.json']
### SOURCE 3: LOGIC & RULES (Knowledge
→ Base): Use the necessary guidelines.
→ [Insert unstructured raw guideline
→ texts]
YOUR TASK: Create a flowchart that
→ logically derives what needs to be
→ done.
Procedure:
1. Look at SOURCE 2: Which topics are
→ active? (e.g., falls, dementia).
2. Go to the appropriate guideline for
→ these topics.
3. Use the data from SOURCE 1 (patient)
→ to check which condition is met.
   - If a condition is NOT met, omit the
     → measure.
Just give me back the Mermaid code.

```

Listing 6: Prompt template for Approach B2. The model must extract logic directly from raw text guidelines without predefined examples.

Evaluation Criteria. The generated Mermaid code was assessed based on:

1. **Syntax Stability:** Whether the generated Mermaid code rendered successfully without manual correction and adhered to valid structural conventions.
2. **Data Fidelity:** Whether patient-specific risks, diagnoses, and contextual constraints were accurately reflected in the generated graph structure.
3. **Logical Coherence:** Whether the generated pathways followed a clinically meaningful sequence without contradictory or structurally inconsistent connections.
4. **Usability:** Whether the resulting graph provided a clear and interpretable overview suitable for rapid situation assessment in nursing workflows.

3.7. Validation Procedure

Both approaches were validated using the dataset of ten synthetic patient profiles. For each scenario, following steps were done:

1. **Execution Matrix:** Prompts were run for Approach A (Image) and Approach B (Code) across all selected models and both languages (English/German).
2. **Syntax Verification (Approach B):** For the code-based approach, it was assessed whether the generated Mermaid.js code rendered successfully without manual correction.
3. **Visual Inspection (Approach A):** For the multimodal approach, the generated dashboards were qualitatively assessed for hallucinations (e.g., legible text vs. gibberish, correct risk color-coding).
4. **Cross-Lingual Check:** The output was compared between German and English prompts to determine if translation layers introduce medical inaccuracies.

3.8. Synthesis and Decision Guidance

Findings are synthesized to identify:

- **Model Efficiency:** Trade-offs between latency-optimized models (e.g., Gemini 3 Fast) and high-reasoning variants (e.g., Gemini 3 Pro/Thinking) for reliable graph generation.
- **Data Robustness:** The comparative reliability of pre-processed logic (JSON) versus flexible full-text extraction.
- **Platform Viability:** The impact of commercial tool constraints (e.g., file limits in Copilot) on clinical workflow integration.
- **Format Suitability:** Guidelines for deploying direct visual generation versus structured Mermaid graphs based on clinical documentation needs.

4. Results

The performance of the generated visualizations was evaluated across ten synthetic patient profiles. The evaluation was carried out by two researchers with completed graduate-level study in the medical and bioinformatics fields. The results are categorized by the generation approach (Multimodal vs. Code-based) and analyzed regarding model capabilities and language robustness.

4.1. Approach A: Visual Fidelity and Semantic Consistency

The direct image generation using NanoBanana Pro produced visually distinct outputs depending on the prompt language and input granularity. While

the requested aesthetic style was generally adhered to by the model, substantial variances in information density and correctness were observed.

Language-Dependent Stability. A comparative analysis of German versus English prompts (using identical JSON inputs) revealed a marked disparity in output quality:

- **German Prompts:** The outputs exhibited sub-optimal spatial organization and high visual noise. The generated dashboards frequently suffered from orthographic instability (hallucinated or misspelled text labels) and displayed a lack of content specificity, often resorting to generic medical imagery rather than patient-specific data points.
- **English Prompts:** These yielded superior structural coherence and orthographic fidelity, with no detected spelling errors. Furthermore, the semantic mapping of risk factors was more precise, evidenced by the correct application of the requested signal color scheme (red/orange) for high-priority pathways.

Complexity vs. Logical Grounding (Full-Text Input). When utilizing Gemini 3 Pro to drive the image generation process based on unstructured full-text guidelines (variation A2 3.5.2), the resulting visualizations increased in graphical complexity. Although the visual output was rendered by NanoBanana Pro, the guiding logic provided by Gemini 3 Pro led to more detailed structures.

- **Superficial Plausibility:** At a macroscopic level, the charts appeared highly detailed and aesthetically professional ("superficial coherence").
- **Semantic Disconnects:** However, a granular inspection revealed severe topological errors, such as dangling edges and erroneous causal links (see Figure 2), indicating a hallucination of logic despite the high visual resolution.

4.2. Approach B: Code-Generated Graph Visualizations

The generated Mermaid.js code was evaluated with respect to visual usability, procedural logic, and information fidelity. The analysis reveals that model architecture and prompt language considerably dictate the utility of the resulting graph, creating a trade-off between overview speed and safety-critical detail.

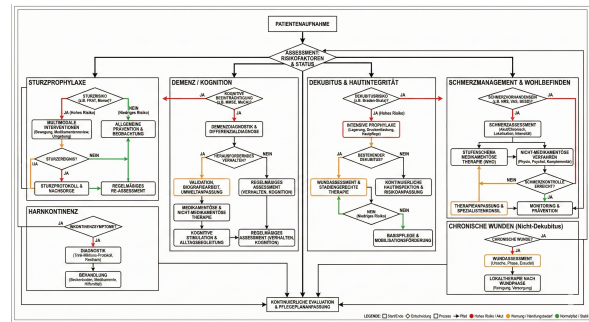


Figure 2: Generated image using Gemini 3 Pro and NanoBanana Pro with German prompt. The output illustrates severe "visual noise" and topological hallucinations, with non-linear pathways and disconnected logical vectors.

4.2.1. Visual Layout and Abstraction

Gemini 3 Fast (German, Variant B1 3.6.1) demonstrated the highest utility for rapid situation assessment. It spontaneously generated styling classes, grouping related interventions into color-coded blocks and omitting inactive pathways. This resulted in a compact "dashboard view" ideal for the point of care.

In contrast, English prompts resulted in unstyled graph topologies with varying degrees of clutter. Gemini 3 Fast (English, Variant B1 3.6.1) exhibited the lowest readability, generating dense cyclic structures with excessive feedback loops (e.g. evaluation arrows returning to the start), creating visual noise. Gemini 3 Pro (English, Variant B1 3.6.1) adopted a structured "Assessment → Action" pattern. While visually clearer than the Fast model, it still listed non-applicable guidelines as isolated nodes. Although it avoided expanding their internal logic, these inactive headers occupied unnecessary visual space, unlike the German output which completely omitted inactive pathways.

However, a notable improvement was observed in Gemini 3 Pro (English, Variant B2 3.6.2). This configuration produced the most aesthetically coherent output of all evaluated permutations, offering a clean, streamlined flow. The primary drawback of this visualization was minor rendering artifacts, where text labels occasionally overlapped with directional arrows, slightly obscuring the path logic (see Figure 3).

4.2.2. Procedural Logic and Precision

Gemini 3 Pro (German, Variant B1 3.6.1) demonstrated the highest procedural fidelity. It adopted a precise "Question → Action" pattern and successfully integrated social resources directly into the clinical workflow, making it the most suitable model for generating concrete nursing instructions.

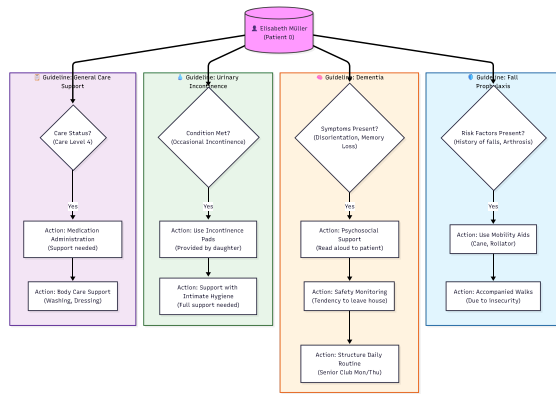


Figure 3: Generated image using Gemini 3 Pro with English prompt and full unstructured text input, showing improved layout but minor text overlap.

In contrast, GPT-5 (Variant B1 3.6.1) produced rigid, list-based structures. Despite the structured input, it listed all theoretically possible measures regardless of specific patient needs, resembling a static checklist rather than a dynamic pathway.

4.2.3. Information Retention and Cross-Lingual Nuances

The qualitative analysis of graph content revealed a substantial interaction between model architecture and prompt language regarding the preservation of psychosocial "soft factors."

- **Narrative Retention vs. Abstraction:** A disparity in the level of detail was observed between the efficiency-focused and high-performance models. Gemini 3 Fast (German, Variants B1 3.6.1 and B2 3.6.2) and GPT-5 demonstrated a tendency toward "terminological abstraction", generalizing specific patient resources (e.g., the calming ritual of reading aloud) into generic nursing categories. In contrast, Gemini 3 Fast (English) maintained higher narrative fidelity, explicitly incorporating the specific action.
- **Model Robustness:** Gemini 3 Pro exhibited the highest degree of information stability across both languages. It successfully mapped the specific biographical resources to the corresponding intervention nodes without loss of detail, regardless of whether the prompt was in German or English. This suggests that higher-parameter models effectively mitigate the abstraction risks inherent in smaller models.
- **Linguistic Stability and Code-Switching:** Despite the high narrative retention in English, Gemini 3 Fast (English, Variants B1

3.6.1 and B2 3.6.2)) showed signs of "translation leakage". While the graph logic was correctly translated, specific technical terms for medical aids remained untranslated within the English node structure ("Promote use of Rollator/Gehstock"). This phenomenon of code-switching indicates a shallower translation layer in efficiency-optimized models compared to the Pro-tier architecture.

4.3. Platform Constraints

A decisive factor in the evaluation was platform accessibility. While the Gemini environment allowed for the seamless processing of the complete guideline set in variation B2 (3.6.2), GPT-5 failed to process the full-text scenario due to strict file upload limits, necessitating the exclusive use of the pre-processed JSON approach for this model.

5. Discussion

This investigation into AI-assisted care pathway visualization highlights trade-offs between aesthetic fidelity, structural verifiability, and preservation of patient-specific context. Analysis of generation errors revealed modality-specific vulnerabilities. In Approach A (image generation), reliance on pixel-based models led to superficial plausibility, with visually coherent outputs often containing topological inconsistencies such as incorrect connections and unsupported clinical elements. In contrast, Approach B (code generation) produced structurally verifiable outputs, as errors typically resulted in rendering failure rather than misleading visuals. However, this approach introduced semantic limitations, as smaller models tended toward terminological abstraction, generalizing individualized interventions and thereby losing important person-centered details. While generative models show promise in reducing nursing cognitive load, direct end-to-end generation remains insufficient for safety-critical applications without robust human oversight.

5.1. The Trap of Superficial Plausibility

A primary finding regarding Approach A (Direct Image Generation) is the discrepancy between visual polish and logical grounding. Models like NanoBanana Pro, even when guided by high-reasoning prompts (Gemini 3 Pro), produced charts with high "superficial plausibility" they visually resembled professional clinical algorithms. However, the presence of "hallucinated logic", such as dangling edges or causally incorrect connections, poses a non-trivial risk. Unlike code-based generation (Approach B), where errors typically manifest as syntax failures (rendering the output blank and

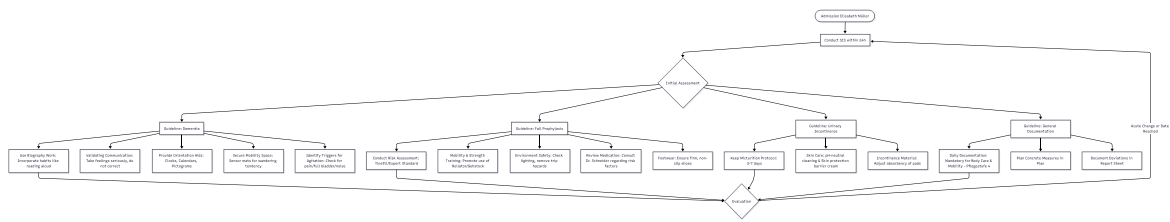


Figure 4: Generated graph code using Gemini 3 Fast with English prompt and structured JSON input.

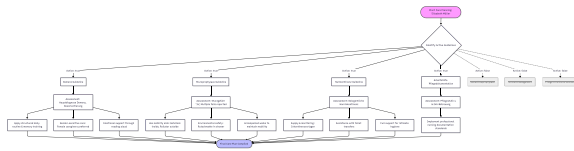


Figure 5: Generated graph code from Gemini 3 Fast with english prompt and unstructured full text input, demonstrating untranslated terminology from the source text.

thus harmless), image generation errors are subtle and integrated into the visual artifact.

5.2. Language-Driven Abstraction vs. Narrative Fidelity

The comparison of prompt languages (English vs. German) uncovered a nuanced trade-off between structural stability and the granularity of care instructions.

- **Syntactic Stability and Code-Switching:** English prompts consistently yielded more stable Mermaid.js code, confirming that models are better attuned to logical control structures in English. However, efficiency-focused models like Gemini 3 Fast exhibited "translation leakage," where technical German terms remained untranslated within the English syntax.
- **Narrative Retention vs. Professional Abstraction:** The results indicate that the native language (German) often triggers a "professional filter" in smaller models. Models like Gemini 3 Fast and GPT-5 tended to abstract specific rituals into generic professional terminology.

This suggests that for clinical applications where the preservation of psychosocial nuances is as critical as medical accuracy, the use of high-performance models (Pro-tier) is mandatory. Efficiency models, while faster, introduce a "granularity risk" where essential person-centered details are sacrificed for either professional abstraction (in native language) or linguistic instability (in English).

6. Conclusion

The evaluation of two AI-assisted approaches for generating patient-specific care pathway visualizations shows that both offer benefits but also notable limitations. Image-based generation provides visually appealing outputs but lacks reliable logical grounding, making it unsuitable for safety-critical contexts. Code-based generation yields structurally verifiable pathways, though results vary with model architecture and prompt language. English prompts favor syntactic stability, whereas German prompts better preserve contextual and psychosocial details.

Overall, while AI-driven visualization can support nursing workflows, neither approach is yet sufficient as an autonomous solution. Future systems should integrate the strengths of both methods, combining visual clarity with semantic and contextual accuracy to more effectively reduce cognitive burden and enhance care efficiency.

7. Outlook

In this setup we tested the use of generated images or graph visualizations without a bigger framework. The results suggest that a useful tool that is easy to use requires more structure and refinement. The tool's output should be as consistent as possible so that nursing staff can familiarize themselves with the flowchart and thus access the information more quickly. This would require a framework which combines machine learning and algorithmic processes.

8. Ethical Considerations

These experiments were conducted using synthetic data, which cannot fully reflect real-world conditions. The synthetic patient data was generated for research purposes only and does not represent real individuals. When real-world data is used, multiple safety measures have to be included. In Europe, for example, such minimum requirements are established and regulated by the European AI Act, the General Data Protection Regulation, and the Medical Devices Regulation. The aim is to protect

individuals' data in the context of patient care. In order to develop a tool to support nursing staff, these regulations must be followed.

9. Limitations

The study also has several limitations:

1. **Synthetic data:** While the generator produces realistic patterns, real clinical records contain more variability and complexity.
2. **Comparability:** The image content created was evaluated visually by hand. No comparative metrics were used.
3. **Sample size:** The size of the patient sets is limited to 10 patients with 7 different care pathways, restricted by the manual effort required for comparability.
4. **No clinical validation:** This study does not include validation and clinical utility with health-care providers.
5. **Model usage:** We used common known n pixel-based image generation models, other models based on logical relationship-based generation might deliver better results. Additionally the used models are paid models. To keep costs low and ensure data security, self-hosted models would also be worth considering for future evaluation.

10. Bibliographical References

- Lemma N. Bulto, Ellen Davies, Janet Kelly, and Jeroen M. Hendriks. 2024. [Patient journey mapping: emerging methods for understanding and improving patient experiences of health systems and services](#). *European Journal of Cardiovascular Nursing*, 23(4):429–433.
- Jean-Baptiste Gartner, Kassim Said Abasse, Frédéric Bergeron, Paolo Landa, Célia Lemaire, and André Côté. 2022. [Definition and conceptualization of the patient-centered care pathway: a proposed integrative framework for consensus](#). *BMC Health Services Research*, 22:558.
- Bonnie Mowinski Jennings, Marianne Baernholdt, and Susan G. Hopkinson. 2022. [Exploring the turbulent nature of nurses' workflow](#). *Nursing Outlook*, 70(3):440–450.
- Matthew Manktelow, Aleeha Iftikhar, Magda Bucholc, Michael McCann, and Maurice O'Kane. 2022. [Clinical and operational insights from data-driven care pathway mapping: a systematic review](#). *BMC Medical Informatics and Decision Making*, 22(1):43.
- Gillian Mould, John Bowers, and M. Ghattas. 2010. [The evolution of the pathway and its role in improving patient care](#). *BMJ Quality & Safety*, 19(5):e14.
- Harrad Rachel and Sulla Francesco. 2018. [Factors associated with and impact of burnout in nursing and residential home care workers for the elderly](#). *Acta Bio Medica : Atenei Parmensis*, 89(Suppl 7):60.
- Jennifer G. Stadler, Kipp Donlon, Jordan D. Siewert, Tessa Franken, and Nathaniel E. Lewis. 2016. [Improving the Efficiency and Ease of Health-care Analysis Through Use of Data Visualization Dashboards](#). *Big Data*, 4(2):129–135.
- Knut Sveidqvist and the Mermaid.js Contributors. 2014. Mermaid.js: Open-Source Diagramming and Charting Tool. <https://mermaid.js.org/>. Accessed: 2026-02-17.
- Xunan Tan, Xiang Suo, Wenjun Li, Lei Bi, and Fangshu Yao. 2025. [Data visualization in healthcare and medicine: a survey](#). *The Visual Computer*, 41(5):3037–3058.
- Britt Wagenaar, Hanneke Merten, and Cordula Wagner. 2026. [Positive and negative experiences and emotions of older patients before, during and after hospital admission: patient journey mapping](#). *BMC Geriatrics*.
- Yu Wang, Ryan A. Rossi, Namyong Park, Nesreen K. Ahmed, Danai Koutra, Franck Dernoncourt, and Tyler Derr. 2025. [Demystifying the Power of Large Language Models in Graph Generation](#). *Findings of the Association for Computational Linguistics: NAACL 2025*, page 8174–8189.
- Yang Wu, Yao Wan, Hongyu Zhang, Yulei Sui, Wucai Wei, Wei Zhao, Guandong Xu, and Hai Jin. 2024. [Automated data visualization from natural language via large language models: An exploratory study](#). *Proceedings of the ACM on Management of Data*, 2(3):1–28.
- Bocheng Zou, Mu Cai, Jianrui Zhang, and Yong Jae Lee. 2024. [VGBench: Evaluating Large Language Models on Vector Graphics Understanding and Generation](#). *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 3647–3659.

A. Generated Graph Data

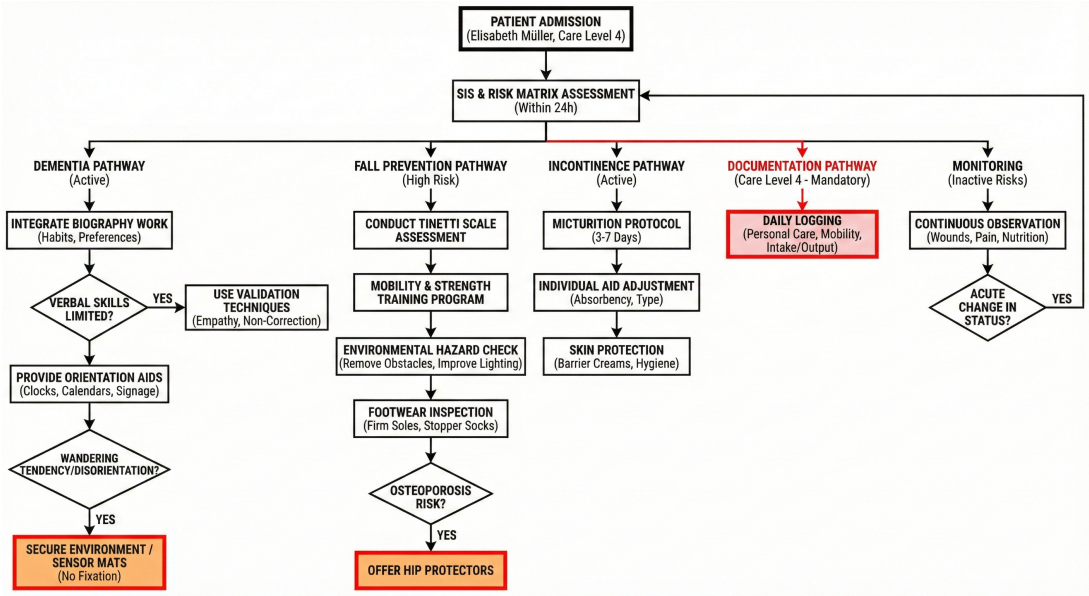


Figure 6: Bigger display of Figure 1, generated image using NanoBanana Pro, with English prompt.

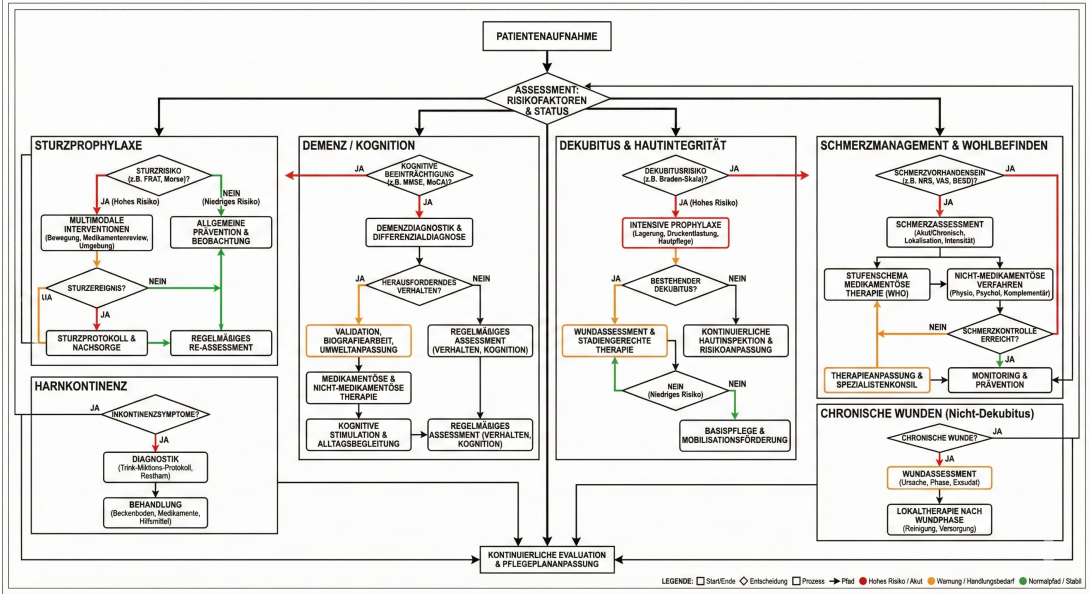


Figure 7: Bigger display of Figure 2, generated image using Gemini 3 Pro and NanoBanana Pro with German prompt.

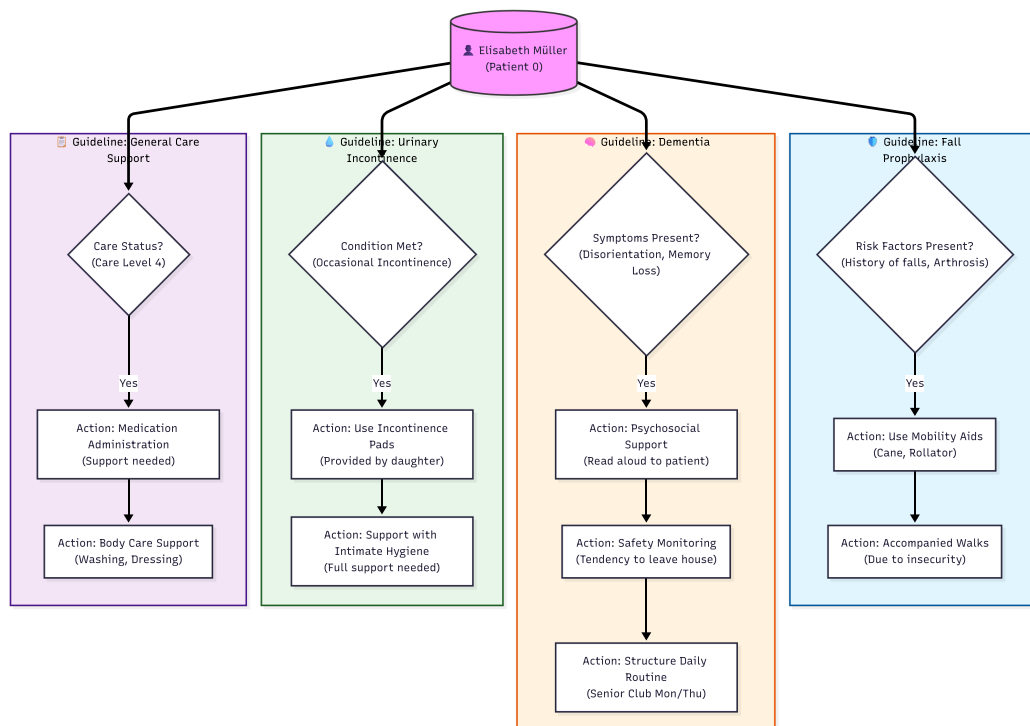


Figure 8: Bigger display of Figure 3, generated image using Gemini 3 Pro with English prompt and full unstructured text input, showing improved layout but minor text overlap.

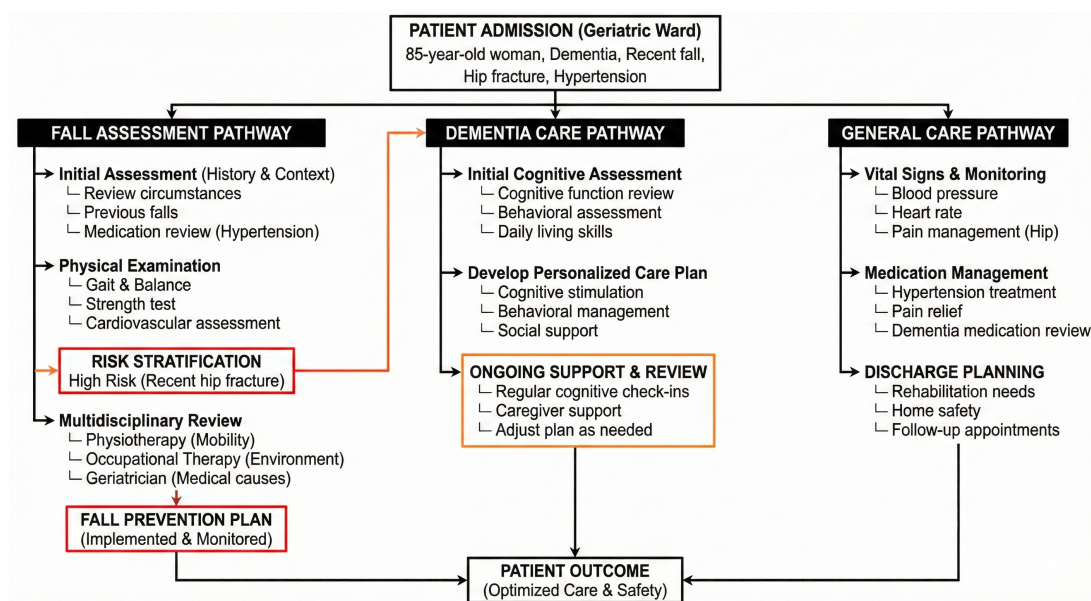


Figure 9: Generated image using Gemini 3 Pro with English prompt and structured guideline input.

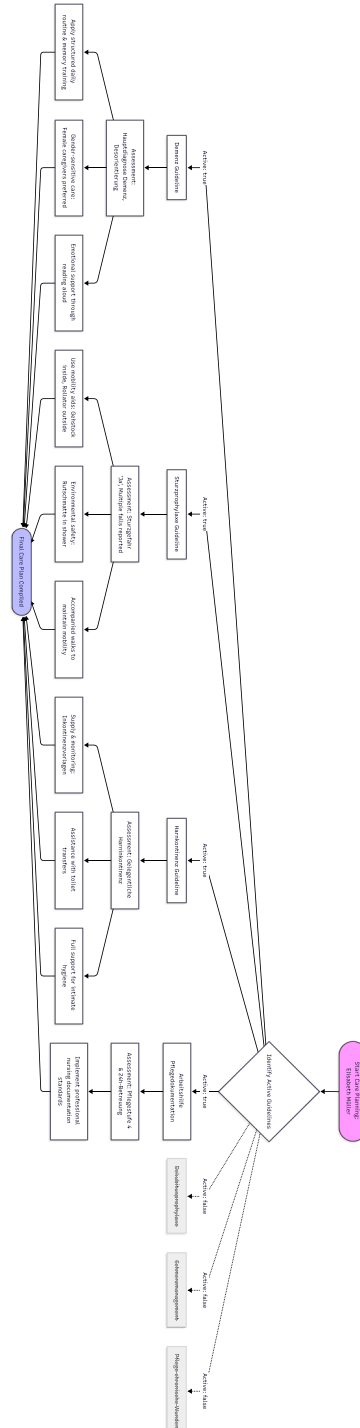
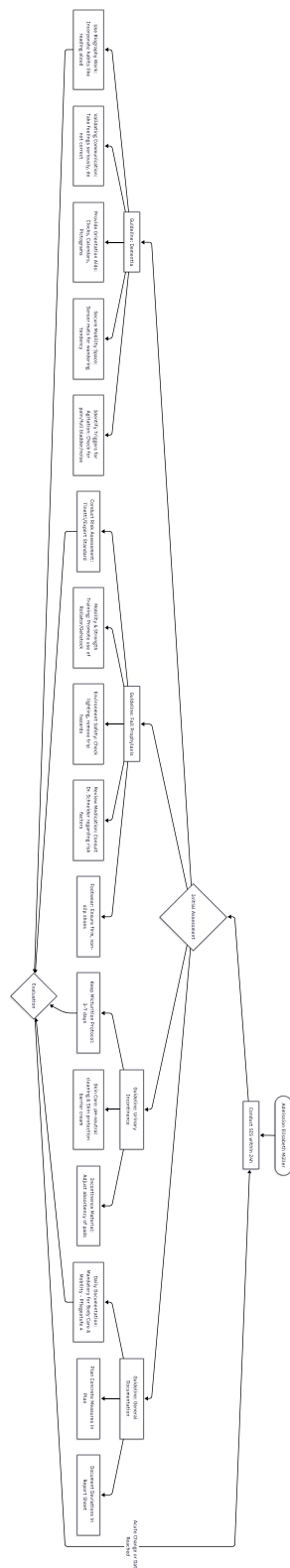
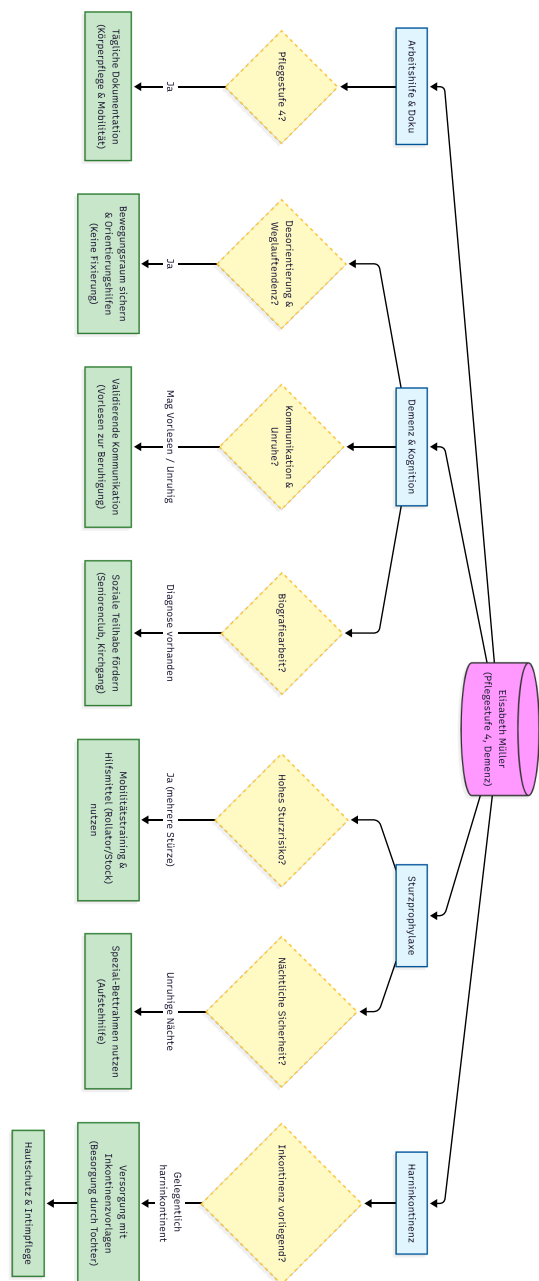


Figure 10: Bigger display of [Figure 4](#) and [Figure 5](#)



(a) Generated graph code using Gemini 3 Fast with German prompt and unstructured full text input.

(b) Generated graph code using Gemini 3 Pro with German prompt and unstructured full text input.

Figure 11: Display of the difference between Gemini 3 Pro and Gemini 3 Fast.

What Makes a Good Doctor Response? A Study on Text-Based Telemedicine

Adrian Cosma[†], Cosmin Dumitrache[‡], Emilian Radoi[‡]

[†]IDSIA, Dalle Molle Institute for Artificial Intelligence, Lugano, CH

[‡]University Politehnica of Bucharest, Bucharest, RO

Correspondence: emilian.radoi@upb.ro

Abstract

Text-based telemedicine has become an increasingly used mode of care, requiring clinicians to deliver medical advice clearly and effectively in writing. As platforms increasingly rely on patient ratings and feedback, clinicians face growing pressure to maintain satisfaction scores, even though these evaluations often reflect communication quality more than clinical accuracy. We analyse patient satisfaction signals in Romanian text-based telemedicine. Using a sample of anonymised text-based telemedicine consultations, we model feedback as a binary outcome, treating thumbs-up responses as positive and grouping negative or absent feedback into the other class. We extract from doctor responses interpretable, predominantly language-agnostic features (e.g., length, structural characteristics, readability proxies), along with Romanian LIWC psycholinguistic features and politeness/hedging markers where available. We train a classifier with a time-based split and perform SHAP-based analyses, which indicate that metadata dominates prediction, functioning as a strong prior, while characteristics of the response text provide a smaller but actionable signal. In subgroup correlation analyses, politeness and hedging are consistently associated with positive patient feedback, whereas lexical diversity shows a negative association.

Keywords: telemedicine; patient feedback; doctor–patient communication; Romanian NLP; interpretability; SHAP; psycholinguistic features.

1. Introduction

Text-based telemedicine has gradually evolved from a niche offering to a more widely used way of receiving care. As more consultations move online, clinicians are increasingly practising medicine in writing. They must give clinically sound advice while explaining it clearly, using language that patients can easily understand and act upon. This change also introduces expectations that have emerged from other online services, especially regarding transparency and feedback. Patients are accustomed to ratings and reviews in other digital services, and many now expect similar indicators when deciding whether to trust a clinician on a telemedicine platform.

As a result, many telemedicine platforms now collect structured feedback after each consultation (e.g., thumbs up / thumbs down or star ratings, optionally accompanied by free-text comments). In contrast to in-person care, where feedback is relatively rare and often informal, online platforms can generate large volumes of explicit evaluations for individual clinicians.

While review systems have become ubiquitous across online platforms, applying them to telemedicine consultations presents unique challenges. Doctors find themselves evaluated not by medical peers, but by patients who generally lack the clinical expertise to assess medical accuracy. This creates a fundamental tension: patient satisfaction and review scores are often driven more by

how medical advice is communicated and framed than by its clinical correctness. Consequently, doctors face pressure to maintain high ratings in order to sustain patient demand, yet they lack clear guidance on which aspects of their communication to change. This creates both an opportunity and a challenge: feedback can incentivize better patient communication and help platforms improve quality, but it can also be difficult for doctors to interpret and act upon, especially when it is negative.

This work attempts to address a key question in telemedicine platforms: *which parts of a doctor's written response are most associated with higher patient satisfaction, and do these patterns differ across patient and clinician subgroups?*

A further motivation for this study is the linguistic context. We analyse interactions written in *Romanian* (Niculae et al., 2025b; Rogoz et al., 2026), a low-resource language (Nigatu et al., 2024). Most *Natural Language Processing* (NLP) research in healthcare has focused on English and other high-resource languages, with relatively limited work on Romanian. As a result, key NLP resources for Romanian, such as clinical-domain models and standardised preprocessing pipelines, remain less developed and less thoroughly validated (Tufiş and Barbu Mititelu, 2014; Dumitrescu et al., 2018). To keep the analysis interpretable and realistic, we therefore rely primarily on language-agnostic, surface-level measurements (e.g., length and structure) and incorporate Romanian-specific resources through the Romanian *Linguistic Inquiry and Word*

Count (LIWC) lexicon (Dudău and Sava, 2022; Pennebaker et al., 2001).

In this work, we use completely anonymised data consisting of Romanian-language text-based telemedicine consultations, to study which factors correlate with patient satisfaction, as reflected in post-consultation ratings. The data underwent irreversible de-identification to ensure that individuals could not be re-identified. Instead of retaining personally identifiable information, individuals were categorized into broad, non-identifiable groups based on general characteristics and prior activity. These groupings include distinctions such as new versus returning users, whether they have previously provided positive feedback or no feedback, and gender. By relying solely on aggregated and non-specific attributes, this approach preserves data utility while maintaining strict privacy protection. We treat feedback as the result of multiple interacting components:

- (i) Patient group information
- (ii) Clinician public profile and platform activity
- (iii) Question metadata (e.g., daytime interval, day of week, chosen medical specialty)
- (iv) Linguistic properties of the doctor response

Importantly, we do not attempt to assess medical correctness, we strictly focus on observable, surface-level attributes of doctors' responses, features that can be measured and acted upon. We also do not include the patient question text in our analysis.

We extract features related to readability and form (e.g., length and structure), psycholinguistic signals (e.g., LIWC categories), and metadata linked to the patient groups, doctors and questions. We then train a gradient-boosting model to predict patient feedback, choosing it as a strong and simple baseline that supports post-hoc interpretability. To understand what the model learns, we use SHAP (Lundberg and Lee, 2017) to quantify global feature importance and provide local explanations for individual interactions, highlighting which response characteristics most strongly influence predicted satisfaction and how these effects vary across patient and doctor subgroups.

2. Related Work

Large language models (LLMs) are increasingly adapted to medical settings via pre-training, fine-tuning, and prompting on clinical corpora (Zhou et al., 2023). Much of the literature evaluates exam-style or multiple-choice QA, where strong performance can be achieved under appropriate adaptation strategies (Zhou et al., 2023). More recent

"copilot" frameworks aim to support broader clinical workflows without task-specific fine-tuning (Ren et al., 2024), but open-ended, real-world interaction settings remain less studied.

Patient-facing deployment depends not only on correctness but also on perceived reliability and empathy. AI-labeled medical advice can be rated as less trustworthy or empathetic, lowering willingness to follow it even when content is held constant (Reis et al., 2024; Hohenstein et al., 2023). Conversely, users judge AI responses as comparable to clinicians despite low accuracy (Shekar et al., 2024). Reliability can also be user-dependent: model behaviour may degrade in public-facing use, plausibly because outputs adapt to the quality and specificity of user inputs (Team; Bean et al., 2025).

Beyond LLMs, medical humanities and sociolinguistics emphasize that patients value empathy, clarity, attention to lived experience and appropriate boundary-setting (Stergiopoulos and Martimianakis, 2023). A well presented medical advice is more likely to have a positive impact (Martin et al., 2005) and to attract positive reviews, directly impacting doctor reputation (Street Jr et al., 2009; Sargeant et al., 2008). Quantitative work links physicians' language to outcomes in online consultations, for example positive-emotion and instrumental language correlating with higher returns (Geng et al., 2024). Classic analyses also report systematic gender differences in doctor-patient dialogue (Meeuwesen et al., 1991), motivating subgroup analyses when studying satisfaction.

Rather than asking whether LLMs provide correct answers, we study what patients reward in human clinician writing. We measure which observable properties of doctor replies (surface form, readability proxies, psycholinguistic features) are associated with positive feedback. This complements Romanian-focused efforts on optimizing LLM-assisted doctor communication (Niculae et al., 2025a). Our goal is not to assess clinical correctness, but to provide an interpretable view of communication cues that correlate with satisfaction.

3. Method

We study anonymised text-based consultations. After an interaction, which consists of a patient question and a doctor response, patients may leave structured feedback (thumbs-up / thumbs-down). In practice most interactions receive no explicit rating and explicit negative feedback is rare. We therefore model patient satisfaction as a binary prediction task, treating thumbs-up as the positive class and grouping thumbs-down and missing feedback into the other class. As such, the task is to predict whether a doctor response will receive positive feedback or not. This formulation roughly corresponds

to the *Positive-Unlabeled learning* (PU-learning) (Bekker and Davis, 2020) setup. The patient and doctor usually do not have a prior connection and each interaction is stand-alone. Patients can either choose which doctor to send their question to, or they can choose to have the doctor assigned by the platform. Although completely anonymised, the data is privately owned and cannot be released.

A particular aspect of the studied text consultations is that doctors may address clarification questions before providing an answer. We retain this structure and distinguish between (i) *Simple Questions* and (ii) *Questions with Clarifications* (a clarification step occurs), which we also use for stratified reporting.

Feature extraction To keep the analysis interpretable and robust in a lower-resource language setting, we focus primarily on surface-level, mostly language-agnostic features, and add Romanian-specific resources where available. We extract features from four sources: (i) patient group information – new vs returning user, prior activity (e.g., previous thumbs-up), (ii) public doctor profile – metadata (e.g., workplace, education), platform activity and historical feedback rate, (iii) question metadata (e.g., daytime interval, day of week, chosen medical specialty), (iv) doctor response text – binned quantities analysed in ranges for length / structure (number of characters, tokens, punctuation such as question marks), concreteness proxies (e.g., number of digits), and lexicon-based markers (e.g., politeness, hedging, recommendation cues), as well as LIWC features. We compute a fixed feature vector, which is used for predictive modelling.

Prediction model and evaluation protocol We train a gradient-boosted decision tree classifier (CatBoost) for binary feedback prediction. To reduce temporal leakage, we use a time-based split: training uses earlier interactions and evaluation is performed on later interactions, for a 80:20 train:test split. We report performance both on the full validation set and on relevant subsets.

Model interpretation and subgroup analyses

To understand which factors drive predictions, we compute SHAP values on the trained model. We summarize global importance with mean absolute SHAP values and use local explanations to inspect individual decisions. In addition, to probe whether associations differ across populations, we perform subgroup correlation analyses between selected text features and feedback, stratifying by patient gender (*patient preferences*) and doctor gender (*doctor behaviour*), and report the resulting patterns descriptively.

Subset	Split	ROC-AUC	PR-AUC	F1	Precision	Recall
Simple Questions	Val	0.6780	0.6865	0.6206	0.6206	0.6209
Only Clarifications	Val	0.7153	0.5857	0.6367	0.6411	0.6493
All Questions	Val	0.7192	0.6313	0.6518	0.6542	0.6577
New Patients	Val	0.6479	0.6003	0.5804	0.6061	0.5963
Returning Patients	Val	0.7656	0.6533	0.6951	0.6968	0.6937
Simple Questions	Val High-conf	0.6951	0.6138	0.6456	0.7724	0.6416
Only Clarifications	Val High-conf	0.8348	0.6542	0.7751	0.7960	0.7603
All Questions	Val High-conf	0.8357	0.6987	0.7889	0.8001	0.7806

Table 1: Feedback prediction performance across question subsets and patient groups (New vs. Returning), reported on the validation split and a high-confidence subset.

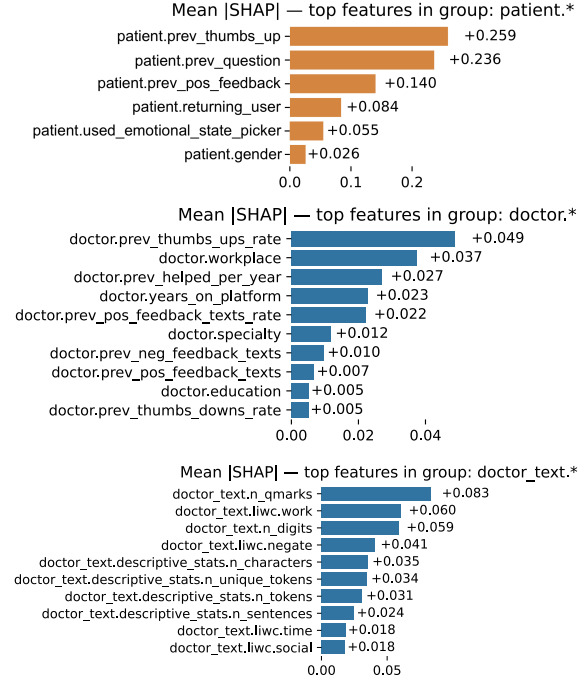


Figure 1: Mean absolute SHAP values for the CatBoost feedback classifier. Larger values indicate greater impact on predictions (magnitude only; SHAP sign is not shown). Quantities are binned and are not processed in absolute numbers.

4. Results

In Table 1, we show classifier performance across subgroups. Since the classifier is properly calibrated (calibration error of 1.5%), high confidence predictions have high performance in terms of ROC-AUC score. Results are also better for clarification-type questions: clarification questions may function as a visible signal of attention and engagement.

Correlation analysis by doctor and patient gender. Figure 2 shows Pearson correlations between doctor-text features and binary feedback, computed separately by doctor gender (*Doctor Behaviour*) and by patient gender (*Patient Preferences*). Across all splits, the largest positive correlations are for politeness markers and hedging in the doctor response text, consistent with prior work (Stergiopoulos and Martimianakis, 2023). Smaller

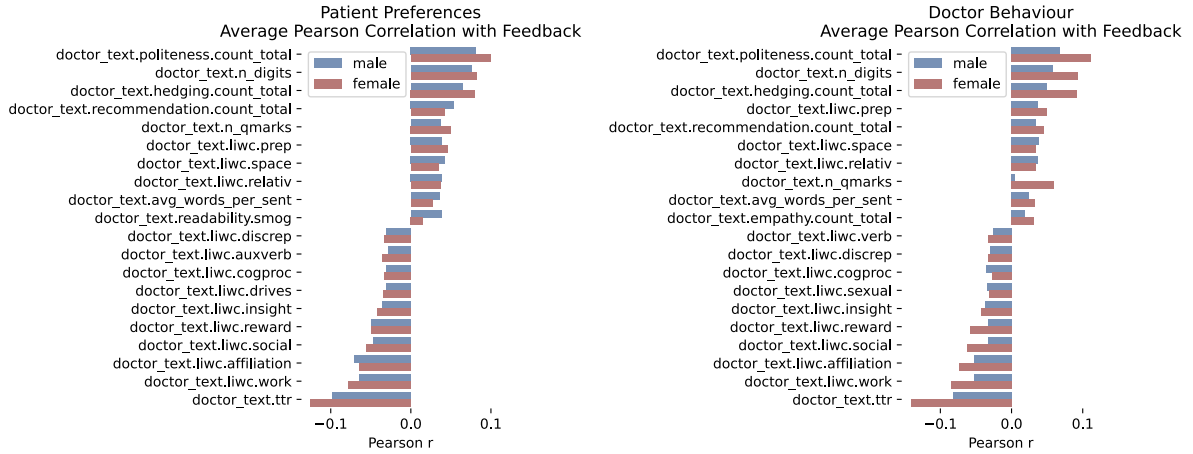


Figure 2: Pearson correlations between selected doctor-response text features and positive feedback, aggregated over all interactions. Positive values indicate features that tend to co-occur with thumbs-up feedback, while negative values indicate features that tend to co-occur with thumbs-down or missing feedback. Quantities are binned and are not processed in absolute numbers.

positive correlations appear for features that often signal concrete, actionable replies, such as the number of digits, explicit recommendations and question marks. The most consistent negative correlation is for lexical diversity, with additional negative correlations for several LIWC categories (e.g., `liwc_work`, `liwc_affiliation`, `liwc_social`). Higher lexical diversity can reflect more complex, less standardized phrasing, which may increase cognitive load; conversely, simpler, more formulaic language may read as clearer and safer (Stergiopoulos and Martimianakis, 2023).

The overall pattern is similar across the two gender groups, suggesting that both male and female patients tend to respond positively to polite, cautious and concrete messages. Differences by patient gender are small. Differences by doctor gender are more visible: for several of the top positive features (especially politeness, hedging and number of digits), correlations are larger within the subset of replies written by female doctors than within those written by male doctors. Correlation magnitudes are modest overall (roughly $|r| \approx 0.02\text{--}0.15$), so these results should be treated as descriptive.

Model explanations using SHAP. Figure 1 reports mean absolute SHAP values (feature importance). Overall, the model relies most on *patient group information* features. The largest contributions come from the returning users feature and the previous thumbs-up feature. These features mainly act as strong priors: patients that previously gave positive feedback are more predictable.

Doctor-level features contribute less than patient group information. The top doctor predictors are previously received thumbs-up rate, doctor workplace, number of previously helped patients, and number of years on the platform, along with past

positive / negative feedback rates. This suggests the model captures persistent differences between clinicians and settings (e.g., workplace) beyond the doctor response.

Doctor response text features have smaller mean $|\text{SHAP}|$ values than patient groups and doctor public profiles features, indicating that metadata dominates the prediction. The most important doctor-text signals include number of questions, number of digits present, and several LIWC categories (e.g., `liwc_work`, `liwc_negemo`, `liwc_time`), as well as length proxies (`n_characters`, `n_tokens`). These patterns are consistent with feedback being associated with replies that are more interactive (questions/clarifications), more concrete (numbers) and more detailed (longer responses), confirming prior work (Geng et al., 2024). Digits often encode dosage, timing, thresholds or step-by-step plans, which are surface markers of actionability. Satisfaction signals are identity- and expectation-loaded, as ratings reflect the rater’s baseline tendency and the clinician’s reputation/context more than the response. This is consistent with “perception-first” evaluation dynamics discussed in prior work (Shekar et al., 2024), and motivates controlling for priors before making communication recommendations.

Figure 3 shows that, when restricting the analysis to doctor-response text and stratifying by medical specialty, the most predictive cues become more specialty-specific. Some features recur across categories, especially digits, time-related language, and task-oriented terms, suggesting that patient feedback favors responses that are concrete, procedural and actionable. At the same time, several categories display clinically intuitive lexical patterns: dermatology emphasizes perceptual

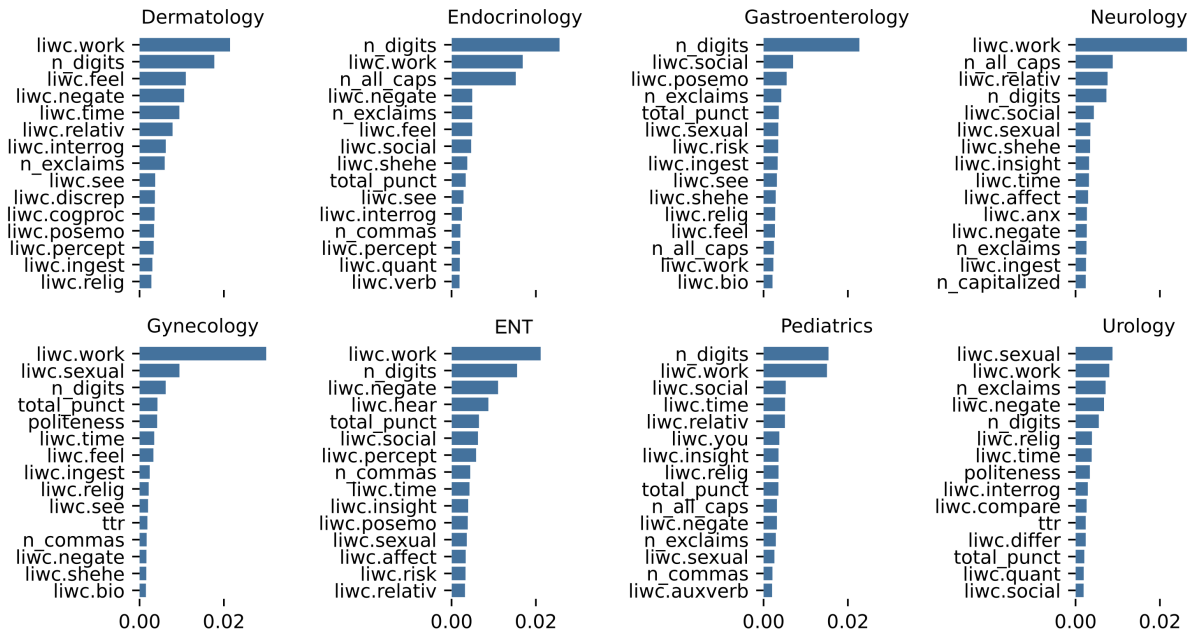


Figure 3: Mean absolute SHAP values for the most important doctor text features, computed separately for each medical question specialty. Bars show the response-text features with the largest contribution magnitude to the feedback prediction model within each specialty. Values indicate feature importance magnitude only and do not show effect direction. Quantities are binned and are not processed in absolute numbers.

and symptom-description language; endocrinology is strongly driven by numeric content; gastroenterology highlights ingestion- and risk-related terms; ENT uniquely surfaces hearing-related language; pediatrics gives more weight to social and second-person terms; and gynecology/urology place greater emphasis on sexual and politeness-related language. Overall, results suggests that positive feedback is associated not only with general clarity, but with responses whose language is appropriately matched to the communicative demands of the medical specialty.

5. Conclusions

We analysed anonymised Romanian text-based telemedicine consultations to examine which characteristics of written clinician replies are associated with positive patient feedback. The main finding is that satisfaction is driven largely by stable priors (patient group information and doctor activity), so predicting feedback mostly captures who tends to rate positively and who tends to receive positive ratings. Still, the response text contains smaller but actionable signals: politeness and hedging are consistently associated with positive feedback, as are concrete, action-oriented cues (e.g., digits and recommendations), while higher lexical diversity correlates negatively. Presumably, more high-level and semantic response features tailored to this spe-

cific use-case could be discovered automatically using feature discovery methods (Cosma et al., 2026). The results suggest that patient feedback should be interpreted primarily as a signal of clear, empathetic written communication, rather than as a proxy for medical accuracy.

Future work should focus on modelling higher-level, semantic features of the doctor responses, such as response clarity and whether the response contains adequate explanations where appropriate. These aspects are not easily captured through lexicon-based approaches and may require the use of pretrained language models (Cosma et al., 2026).

6. Bibliographical References

- Andrew Bean, Rebecca Elizabeth Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera-Gómez, Sara Hincapié, Aruna Ekanayaka, Lionel Tarassenko, Luc Rocher, et al. 2025. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*.
- Jessa Bekker and Jesse Davis. 2020. Learning from positive and unlabeled data: A survey. *Machine learning*, 109(4):719–760.

- Adrian Cosma, Oleg Szehr, David Kletz, Alessandro Antonucci, and Olivier Pelletier. 2026. [Automatic Prompt Optimization for Dataset-Level Feature Discovery](#).
- Diana Paula Dudău and Florin Alin Sava. 2022. The development and validation of the Romanian version of Linguistic Inquiry and Word Count 2015 (Ro-LIWC2015). *Current Psychology*, 41(6):3597–3614.
- Stefan Daniel Dumitrescu, Andrei Marius Avram, Luciana Morogan, and Stefan-Adrian Toma. 2018. RoWordNet—A Python API for the Romanian WordNet. In *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–6. IEEE.
- Shuang Geng, Yuqin He, Liezhen Duan, Chen Yang, Xusheng Wu, Gemin Liang, and Ben Niu. 2024. The association between linguistic characteristics of physicians’ communication and their economic returns: mixed method study. *Journal of Medical Internet Research*, 26:e42850.
- Jess Hohenstein, Rene F Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F Jung. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific reports*, 13(1):5487.
- Scott M. Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). *CoRR*, abs/1705.07874.
- Leslie R Martin, Summer L Williams, Kelly B Haskard, and M Robin DiMatteo. 2005. The challenge of patient adherence. *Therapeutics and clinical risk management*, 1(3):189–199.
- Ludwien Meeuwesen, Cas Schaap, and Cees Van Der Staak. 1991. Verbal analysis of doctor-patient communication. *Social science & medicine*, 32(10):1143–1150.
- Andrei Niculae, Adrian Cosma, Cosmin Dumitrache, and Emilian Radoi. 2025a. [Dr. Copilot: A Multi-Agent Prompt Optimized Assistant for Improving Patient-Doctor Communication in Romanian](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1780–1792, Suzhou (China). Association for Computational Linguistics.
- Andrei Niculae, Adrian Cosma, and Emilian Radoi. 2025b. [A Retrieval-Based Approach to Medical Procedure Matching in Romanian](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 167–175, Viena, Austria. Association for Computational Linguistics.
- Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. [The Zeno’s Paradox of ‘Low-Resource’ Languages](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.
- James W Pennebaker, Martha E Francis, Roger J Booth, et al. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.
- Moritz Reis, Florian Reis, and Wilfried Kunde. 2024. Influence of believed AI involvement on the perception of digital medical advice. *Nature Medicine*, 30(11):3098–3100.
- Zhiyao Ren, Yibing Zhan, Baosheng Yu, Liang Ding, and Dacheng Tao. 2024. Healthcare copilot: Eliciting the power of general llms for medical consultation. *arXiv preprint arXiv:2402.13408*.
- Ana-Cristina Rogoz, Radu Tudor Ionescu, Alexandra-Valentina Anghel, Ionuț-Lucian Antone-lordache, Simona Coniac, and Andreea Iuliana Ionescu. 2026. [A large-scale benchmark for evaluating large language models on medical question answering in Romanian](#). *npj Digital Medicine*.
- Joan Sargeant, Karen Mann, Douglas Sinclair, Cees Van der Vleuten, and Job Metsemakers. 2008. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Advances in Health Sciences Education*, 13:275–288.
- Shruthi Shekar, Pat Pataranutaporn, Chethan Sarabu, Guillermo A Cecchi, and Pattie Maes. 2024. People over trust AI-generated medical responses and view them to be as valid as doctors, despite low accuracy. *arXiv preprint arXiv:2408.15266*.
- Erene Stergiopoulos and Maria Athina Tina Martimianakis. 2023. What makes a ‘good doctor’? A critical discourse analysis of perspectives from medical students with lived experience as patients. *Medical Humanities*, 49(4):613–622.
- Richard L Street Jr, Gregory Makoul, Neeraj K Arora, and Ronald M Epstein. 2009. How does communication heal? Pathways linking clinician–patient communication to health outcomes. *Patient education and counseling*, 74(3):295–301.
- Anthropic Team. [Anthropic economic index report: Economic primitives](#). Accessed: 18 Feb. 2026.

Dan Tufiş and Verginica Barbu Mititelu. 2014. *The Lexical Ontology for Romanian*, volume 48 of *Text, Speech and Language Technology*, pages 491–504. Springer.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jing Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

Datasets for a Chatbot for Clinical Trial Search

Yumeng Yang, MS¹, Ethan B Ludmir, MD², Kirk Roberts, PhD¹

¹McWilliams School of Biomedical Informatics

The University of Texas Health Science Center at Houston, Houston, TX, USA,

²Department of Radiation Oncology

The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Yumeng.Yang@uth.tmc.edu

Abstract

Matching patients to clinical trials is a critical bottleneck hindered by complex eligibility criteria. While conversational AI offers a promising solution, its safe deployment depends on high-quality, domain specific data. This paper introduces three benchmark datasets designed to support the development and evaluation of conversational agents for clinical trial pre-screening. First, a manually-annotated paired-criterion dataset provides a gold standard for structuring raw criteria, which we used to objectively group 12,596 criteria. Second, we curated a human-authored question benchmark to validate the clinical fidelity and patient-centric clarity of questions generated by a medical LLM, ensuring the AI's dialogue is accurate and understandable. Third, we constructed a human-validated assessment corpus of criterion-question-answer tuples with human-labeled outcomes to evaluate criterion classification based on a patient's answer to a generated question. The primary contribution of this work is a foundational set of benchmark datasets, designed to support and evaluate key components for a chatbot for clinical trial search.

Keywords: Eligibility Criteria, Natural Language Processing, Large Language Models, Conversational AI

1. Introduction

The matching of eligible patients to suitable clinical trials is a critical challenge in finding treatments for diseases such as cancer (Unger et al., 2016; Murthy et al., 2004). While clinical trials offer access to potentially life-saving innovative treatments, patient enrollment rates remain persistently low, with estimates suggesting that fewer than 5% of adult cancer patients participate (Unger et al., 2016; Murthy et al., 2004). A barrier to enrollment is the complexity of clinical trial eligibility criteria (Kim et al., 2017; Idnay et al., 2021), which are often articulated in dense, technical prose across disparate trial documents. This information overload makes it exceedingly difficult for patients and even clinicians to efficiently identify appropriate trials, leading to missed opportunities and delays in care.

The conventional approach to trial recruitment relies on manual eligibility screening by clinical staff, a process that is notoriously laborious and time-consuming (Cowie et al., 2017; Köpcke and Prokosch, 2014). Additionally, patients often lack effective tools to actively find potentially suitable clinical trials (Borno et al., 2022). Recent advances in natural language processing (NLP) and large language models (LLMs) present a promising avenue for mitigating this challenge. Conversational AI systems, in particular, have the potential to democratize access to trial information by transforming the complex pre-screening process into an intuitive, interactive dialogue. Such systems can guide patients through a series of clarifying questions to determine their potential eligibility. However, the

efficacy, safety, and reliability of these clinical AI tools are not only a function of model architecture; they are fundamentally dependent on the quality of the underlying data resources used for their development and validation.

This paper addresses a gap between the potential of conversational AI in clinical trial matching and the practical, resource-intensive work required to build a trustworthy system. We argue that the development of robust language resources, accompanied by a rigorous, multi-stage evaluation protocol, is a prerequisite for a successful clinical application and constitutes a significant research contribution in itself. We present a detailed case study of our data preparation, annotation, and evaluation pipeline designed to support a conversational AI for cancer trial pre-screening. We specifically focus on cancer clinical trials, but none of our methods are limited to cancer; it is just the most prominent use case for clinical trials. Our work is organized around three data-centric stages: (1) clustering of semantically similar eligibility criteria, (2) generation of natural language questions for each criteria cluster, and (3) assessment of patient answers against the source criterion.

Further, we evaluate several state-of-the-art methods as baselines for creating and validating the linguistic assets at each stage. For criteria clustering, we systematically compare the performance of a density-based algorithm, HDBSCAN (Campello et al., 2013), against the centroid-based K-Means, establishing a more effective method for grouping nuanced clinical statements. For question generation, we develop a gold-standard dataset

of human-authored questions and use it to benchmark and iteratively refine questions generated by a specialized medical LLM, MedGemma (Søllergren et al., 2025). This iterative refinement process, guided by metrics such as BERTScore (Zhang et al., 2020) and ROUGE-L (Lin, 2004), resulted in a highly optimized prompt engineering strategy. Finally, we detail the creation of a human-validated assessment corpus to evaluate the LLM’s ability to accurately map a user’s answers to the underlying eligibility criterion, ensuring the final system’s reasoning is clinically sound.

The primary contribution of this work is not the conversational agent itself, but the datasets that enable its construction. We envision a variety of conversational architectures are possible. However, the three sub-tasks studied here are fundamental to developing interactive systems for clinical trials. These tasks represent the basic building blocks of (i) reducing the dimensionality of clinical trial eligibility criteria, (ii) converting the criteria to a conversational prompt, and (iii) filtering trials based on the user’s responses. In addition, this paper provides the initial prompts used to generate questions and assess patient responses against eligibility criteria, which can serve as a baseline for future refinement.

2. Related Work

Early efforts to automate patient-trial matching primarily relied on structured data and rule-based systems. Researchers developed systems that converted free-text eligibility criterion into structured representations using medical ontologies like UMLS and SNOMED-CT, enabling queries against patient data in Electronic Health Records (EHRs) (Weng et al., 2011; Yuan et al.). These approaches were often labor-intensive, requiring extensive manual effort to create and maintain rules, and struggled to capture the semantic nuance inherent in clinical prose (Weng et al., 2011). The emergence of machine learning has led to the development of methods include information extraction and text classification. These systems used classic Natural Language Processing (NLP) pipelines, including Named Entity Recognition (NER) and relation extraction, to identify clinical concepts within the trial criteria (Kang et al., 2017). However, these models were dependent on feature engineering and large, domain-specific annotated datasets, which are expensive and time-consuming to create. The paradigm shifted significantly with the development of deep learning-based contextual embeddings. Domain-specific pre-trained models such as ClinicalBERT (Alsentzer et al., 2019) and BioBERT (Lee et al., 2020), which are fine-tuned on clinical and biomedical corpora, demonstrated a superior

ability to understand the semantics of medical text. These models have become the backbone for numerous downstream clinical tasks, including cohort identification and criteria analysis.

More recently, generative LLMs like Med-PaLM (Singhal et al., 2023) and Med-Gemini (Saab et al., 2024) have shown remarkable capabilities in complex clinical reasoning and natural language generation. Research has showcased their potential for tasks like medical question answering and clinical note summarization (Nori et al., 2023). However, a significant portion of this research focuses on end-task performance, often overlooking the critical, resource-intensive data preparation pipeline required for safe and reliable deployment. Our previous work includes manually annotating whether a cancer trial includes or excludes a key criterion in its protocol, which is important for downstream tasks (Yang et al., 2024a,b). This highlights a crucial gap between demonstrating a model’s potential and engineering a trustworthy clinical application.

Our work’s primary contribution is the methodology for building the data foundation for a clinical trial conversational agent. This involves three key NLP tasks: structuring the source criteria, generating patient-facing questions, and classifying the user’s response. Clustering has been explored for discovering topics in clinical text, such as grouping patient notes or research articles (Arnold et al., 2016). However, to our knowledge, an evaluation of different embedding strategies and clustering algorithms specifically for the nuanced text of eligibility criteria has been lacking. Our analysis of SBERT vs. other embeddings and K-Means vs. HDBSCAN provides a methodological contribution to this niche area. Furthermore, while question generation is an established NLP task, its application to high-stakes clinical scenarios necessitates a focus on quality and fidelity. The standard practice of relying solely on automated metrics like ROUGE or BERTScore (Zhang et al., 2020) is insufficient where patient safety is concerned. This paper, therefore, outlines a replicable, human-in-loop methodology for generating and assessing conversational datasets to better support the clinical trial screening process.

3. Experimental Setup

This paper presents a multi-stage pipeline to construct and validate the linguistic resources for a conversational AI system for trial pre-screening. The methodology is partitioned into three stages: (1) the unsupervised clustering of eligibility criteria to distill recurring clinical concepts; (2) the generation and evaluation of natural language questions for each cluster; and (3) the development of a validated corpus to assess a large language model’s ability to accurately interpret patient responses.

Medical Concepts (Cluster)	Raw Eligibility Criteria
Dihydropyrimidine Dehydrogenase (DPD) Deficiency	1. Known dihydropyrimidine dehydrogenase (DPD) deficiency. 2. Known DPD deficiency tested by measuring uracil level in blood or by genetic testing for DPD mutations.
Eastern Cooperative Oncology Group (ECOG) Performance Status ≤ 1	1. Eastern Cooperative Oncology Group (ECOG) performance status of 0 or 1. 2. Patients must have an ECOG performance status score of ≤ 1.

Table 1: Example eligibility criteria for Dihydropyrimidine Dehydrogenase (DPD) Deficiency and Eastern Cooperative Oncology Group (ECOG) Performance Status ≤ 1 .

3.1. Clustering of Eligibility Criteria

The initial stage focused on structuring the raw criteria text into semantically coherent groups. To begin, a balanced dataset was curated by randomly sampling 100 clinical trials for each of five major cancer types—breast, lung, prostate, colorectal, and pancreatic cancer—resulting in a corpus of 500 trials. All inclusion and exclusion criteria were extracted from these trials, yielding 12,596 individual criteria. To prepare this text for clustering, each criterion was transformed into a high-dimensional vector using four distinct embedding strategies: the contextually-aware Sentence-BERT (SBERT) and Universal Sentence Encoder (USE), the domain-adapted ClinicalTrialBERT (Yang et al., 2024b), and a classical lexical baseline using Term Frequency–Inverse Document Frequency (TF-IDF) with dimensionality reduction via Singular Value Decomposition (SVD). These embedded representations were subsequently grouped using two complementary clustering algorithms: the centroid-based K-Means, configured with $k=50$ and $k=300$ to capture varying granularities, and the density-based HDBSCAN, tested under tight, moderate, and relaxed density parameterizations. This approach produced an ensemble of 20 unique clustering configurations for evaluation.

To identify the optimal clustering configuration, we manually annotated a set of 500 criterion pairs selected from the most frequent disagreements among our 30 model variants. These pairs were identified by calculating their co-clustering frequency and sampling from pairs that were grouped

together in 5 to 15 of the 20 runs. Two annotators labeled these pairs based on whether they represented the same clinical concept and could be addressed by a single question, achieving a strong inter-annotator agreement with a Cohen’s Kappa of 0.742, and agreement of 0.915.

Each of the 20 configurations was then benchmarked against this gold standard, with performance assessed using precision, recall, and F1. The results indicated that SBERT embeddings provided the most effective representations, achieving the highest average F1 (0.71) across all clustering methods. While HDBSCAN configurations consistently yielded high precision at the cost of lower recall, K-Means offered a more balanced performance. The best configuration, SBERT combined with K-Means ($k=200$), achieved a superior F1 of 0.84 (precision=0.94, recall=0.76). Consequently, this model was selected to generate the definitive criteria clusters for all subsequent stages of the pipeline. The detailed results for each embedding method are listed in Table 2.

While K-means was used to help build the gold standard, it is not included in our selection of clustering algorithms because it requires the number of clusters to be specified a priori, which depends heavily on the specific disease and number of trials. Instead, our clustering experiments for an algorithm to actually include in the chatbot relies on HDBSCAN, which determines the optimal number of clusters adaptively based on the intrinsic structure of the data. To determine the optimal hyperparameters for HDBSCAN clustering on SBERT-encoded eligibility criteria, we conducted a systematic grid search over a broad parameter space. The four key parameters controlling cluster granularity were: minimum cluster size ($\text{min_cluster_size} \in \{5, 20, 50, 100\}$), minimum samples ($\text{min_samples} \in \{1, 2, 3, 4, 5\}$), cluster selection epsilon ($\epsilon \in \{0.3, 0.5, 0.6, 0.7, 0.8, 0.9\}$), and alpha ($\alpha \in \{0.5, 0.7, 0.9, 1.0\}$). For each configuration, clustering performance was quantified in terms of precision, recall, F1, and accuracy, where a pair was considered correctly predicted if both sentences appeared in the same cluster.

The grid search results demonstrated that clustering performance was relatively stable across a narrow range of high-performing configurations ($F1 = 0.91$). Configurations with smaller min_cluster_size and min_samples values tended to yield higher recall and finer-grained clusters, whereas larger values improved precision but produced coarser groupings. Reducing ϵ below 0.8 increased the total number of clusters but did not substantially affect F1 performance, suggesting that clustering quality was robust to moderate density threshold variations.

Based on this analysis, we selected the config-

Embedding	Avg Accuracy	Avg Precision	Avg Recall	Avg F1
ClinicalTrialBERT	0.5096	0.9182	0.4652	0.5668
SBERT	0.6508	0.9436	0.6326	0.7131
TFIDF	0.5384	0.9013	0.5132	0.5949
USE	0.5892	0.9182	0.5727	0.6505

Table 2: Evaluation of Embedding Models

uration of HDBSCAN with `min_cluster_size = 5`, `min_samples = 1`, $\varepsilon = 0.7$, and $\alpha = 0.9$. This configuration achieved a balanced trade-off (precision = 0.841, recall = 0.995, F1 = 0.911), while producing a reasonable number of clusters ($n = 425$). This balance provided both high intra-cluster homogeneity and sufficient granularity for downstream question generation and patient-trial matching tasks.

3.2. Question Generation and Evaluation

Following the identification of semantically coherent eligibility criteria clusters, the next stage focused on generating a representative, patient-facing question for each cluster. This benchmark dataset serves as a standardized foundation for evaluating conversational agents in clinical trial screening, ensuring consistency and reproducibility across different models and cancer types. By creating representative questions for each cluster, we capture common semantic patterns in eligibility criteria and translate them into language that patients can easily understand, thereby bridging the gap between clinical and layperson terminology.

Asking the right questions is essential for efficiently filtering ineligible trials and improving screening precision. For example, instead of vague, model-generated questions like “How would you describe your ability to carry out daily activities?”, it is more effective to ask directly about the patient’s ECOG performance status, which preserves clinical specificity and avoids ambiguity. For criteria such as “ECOG between 0–2,” the model was instructed to generate open-ended rather than binary questions, enabling downstream assessment modules to interpret responses more flexibly and accurately. This benchmark data set supports the evaluation of question generation models and provides a reusable framework to evaluate conversational AI systems in healthcare contexts.

We utilized a domain-specific large language model, MedGemma 27b, within a prompt engineering framework to generate these questions. The prompt’s design (shown in Table 3) was guided by several core principles established through iterative refinement. The primary goal was to generate a single, specific question whose answer would be directly classifiable (e.g., as a number or a yes/no response). Key instructions emphasized preserving all medical specificity, directing the model to retain all technical terms, units, and test names

(like ECOG, ALT, or g/dL) rather than generalizing them. Furthermore, the prompt explicitly guided the model to prioritize quantifiable answers—asking for a patient’s specific lab value or score instead of a simple binary confirmation—and to translate complex terms by adding brief, parenthetical clarifications (e.g., “ECOG (Eastern Cooperative Oncology Group)”). This structured approach, which also included rules for handling logical conditions (e.g., ‘OR’ statements), was engineered to produce questions that were clinically precise, unambiguous for a layperson, and machine-readable for downstream assessment. To validate their quality, the machine-generated questions were benchmarked against a gold-standard set of human-made questions. The evaluation relied on BERTScore for measuring semantic similarity, thereby quantifying the fidelity of the generated questions to the human references. We manually selected 100 high-quality clusters for annotation, which were defined by two criteria: first, the cluster had a clear medical theme, with most criteria pertaining to the same concept (e.g., Eastern Cooperative Oncology Group (ECOG)); and second, the majority of criteria within a cluster could be consolidated and addressed by a single question, despite being expressed differently.

The benchmark dataset was created by two annotators who made questions for each of the 100 clusters. We observed that some clusters contain multiple themes; in these cases, we generated multiple questions to ensure complete coverage. Following a consolidation phase between the two annotators, the final corpus served as our gold-standard benchmark. For the experiment, the model was constrained to generate a single question per cluster, which was then evaluated for semantic fidelity against the human-made references for that cluster using BERTScore. The evaluation yielded a mean precision of 0.3256, mean recall of 0.5278, and a mean F1 of 0.4234, indicating a moderate level of semantic alignment between the model-generated and human-authored questions. These results suggest that while the model captures key aspects of the reference questions, there remains room for improvement in fully mirroring human intent and phrasing. To better illustrate the effectiveness of our approach, several representative examples from the benchmark are shown below. For instance, within the functional status cluster, the criterion “ECOG between 0–2” led to the question

Prompt Type	Prompt Content
Question Generation Prompt	GOAL: Analyze each clinical trial eligibility criterion and generate a single, specific question whose answer can be directly used to determine eligibility (classifiable as yes/no or numeric). CORE PRINCIPLES: 1. Preserve all medical specificity – keep every named medical entity, test, score, or unit (e.g., ECOG, KPS, WHO, albumin, hemoglobin, ALT, AST, creatinine). Do not generalize or replace these with vague phrases. 2. Prioritize quantifiable or discrete answers – if the criterion includes a number, grade, or threshold (e.g., ECOG ≤ 1, Albumin ≥ 3.0 g/dL), ask for that number explicitly and encourage patients to find it in their medical record. 3. Translate, don't simplify away meaning – keep technical terms and add short clarifications when needed (e.g., ECOG performance status → ECOG (Eastern Cooperative Oncology Group) performance status). 4. Handle logical conditions explicitly – include both sides of logical conditions (e.g., bleeding disorders OR taking anticoagulants) in the question. 5. Be clinically actionable – phrase questions so clinicians can use answers directly for eligibility screening. EXAMPLES: • ECOG performance status of 0 or 1 → Q: What is your most recent ECOG (Eastern Cooperative Oncology Group) performance status score? It is a number from 0 to 5 in your oncology record. • Serum albumin ≥ 3.0 g/dL → Q: What is your most recent serum albumin level, as listed in your blood test results? • History of bleeding disorders OR current use of blood thinners → Q: Do you have a bleeding disorder or are you currently taking blood-thinning medication such as warfarin or apixaban? OUTPUT FORMAT: Generated Question: "[your question here]" Criterion: {rep_criteria}
Assessment Prompt	You are assessing patient eligibility against a clinical trial criterion. Inputs: Patient Answer: "{patient_answer}" Question: "{question}" Criterion Type: {criterion_type} Criterion: {criterion_text} Instructions: • If INCLUSION criterion: patient must meet this requirement. • If EXCLUSION criterion: patient must not meet this requirement. Decision options (respond with exactly one): • INCLUDE – patient meets inclusion or does not meet exclusion. • EXCLUDE – patient does not meet inclusion or meets exclusion. • UNKNOWN – insufficient data to determine. Output: provide only one decision word (INCLUDE / EXCLUDE / UNKNOWN), nothing else.

Table 3: Prompts for question generation and assessment

“What is your ECOG score?”, which promotes more flexible downstream assessment than a binary version. Similarly, in the hematologic function cluster (e.g., “Absolute neutrophil count $\geq 1500/\text{mm}^3$ ”), the generated question “Do you know your most recent blood test results, including neutrophil or platelet counts?” successfully balances clinical specificity and patient readability. Finally, for organ function and cancer stage clusters, questions such as “Have you been told that your kidney function is normal?” and “What type and stage of cancer were you diagnosed with?” demonstrate how technical eligibility language can be converted into patient-centered prompts without loss of medical meaning.

3.3. Patient Answer Assessment and Validation

The third stage established a framework to evaluate the conversational agent’s ability to assess patient answers and make eligibility decisions. This benchmark dataset enables end-to-end assessment of the clinical trial screening pipeline—from question generation to response interpretation and eligibility assessment.

To establish a framework for evaluating conversational eligibility assessment, we developed a corpus of 13,344 labeled tuples, each comprising a criterion, a corresponding patient-facing question, and responses (answers). The dataset integrates two distinct components: a machine-generated set of 6,660 examples, where the model was prompted to produce two opposing synthetic answers per criterion, and a human-generated answers with set of 6,684 examples, where real users provided up to three nuanced responses per question. To ensure these answers reflected real-world screening scenarios, we intentionally included a spectrum of response types. Beyond straightforward “yes” and “no” answers, the dataset incorporates conditional responses (e.g., “I can stop my medication if required”) and “incomplete” entries where patients cannot recall specific dates or lab values. This design forces the model to move beyond binary logic and handle cases requiring follow-up or manual review.

Evaluating the MedGemma 27B model against this human-labeled gold standard revealed a three-class accuracy of 74.1% and a binary (non-exclusion vs. exclusion) accuracy of 81.3%. The model demonstrated reliable recognition of the INCLUDE class (precision = 0.83, recall = 0.72, F1 = 0.77) and the EXCLUDE class (precision = 0.69, recall = 0.82, F1 = 0.75). However, the UNKNOWN class exhibited a strong dependence on label prevalence; in the machine-generated subset where UNKNOWN frequency was low (5.7%), the model struggled significantly (F1 = 0.32), whereas per-

formance improved in the human-annotated subset (19.8% prevalence) to an F1 of 0.68. Across the aggregate dataset, the UNKNOWN class achieved an overall F1 of 0.61. These results indicate that while the model effectively identifies explicit affirmations and denials, its ability to defer for clarification is sensitive to data distribution and the inherent ambiguity of natural patient language.

The failure cases listed in Table 5 underscore the critical need for our assessment dataset. While large language models demonstrate powerful general reasoning capabilities, their assessments varied when faced with domain-specific challenges such as unit conversion and contextual interpretation. This highlights that careful, targeted evaluation is essential for specialized applications. Our dataset, therefore, provides a crucial benchmark for identifying these gaps and driving the development of more robust and reliable conversational AI for assessing clinical trial eligibility.

4. Results

Our multi-stage data generation pipeline produced three validated benchmark datasets designed to support and evaluate the development of conversational AI for clinical trial screening. The first dataset is a manually annotated gold standard of paired eligibility criteria, which played a critical role in objectively determining the optimal clustering configuration. By evaluating 20 different configurations against this benchmark, we identified SentenceBERT (SBERT) as the superior embedding model, achieving the highest average F1-score of 0.7131. The final optimized model, combining SBERT with HDBSCAN, demonstrated high performance with an F1 of 0.911 (precision = 0.841, recall = 0.995) on our benchmark. The application of this validated model successfully structured 12,596 raw eligibility criteria into a set of 425 clusters.

The second benchmark data is the human-made question dataset corresponding to 100 criteria clusters. To ensure comprehensive coverage, our annotators manually crafted multiple valid, patient-facing questions for clusters that include more than one clinical theme. This design provides a flexible evaluation target; an automated system is considered successful if it generates a question semantically similar to any of the human-made variants for a given cluster. We used this benchmark to evaluate the questions generated by the MedGemma model, using BERTScore to measure semantic fidelity. The evaluation yielded a mean precision of 0.3256, mean recall of 0.5278, and a mean F1 of 0.4234. This quantitative result indicates a moderate level of semantic alignment, confirming that while the model captures key aspects of the clinical intent (as shown in Table 4), there remains signifi-

Cluster Theme		Human-generated Question	Machine-generated Question
Functional	Status (ECOG)	What is your Eastern Cooperative Oncology Group (ECOG) score?	What is your most recent Eastern Cooperative Oncology Group (ECOG) score?
Liver Function (AST)		Any recent blood tests or scans your doctor or nurse mentioned?	What are your most recent AST (aspartate aminotransferase) or ALT (alanine aminotransferase) levels?
Renal	Function (Urine Protein)	Could you describe your most recent urine test results?	What is your most recent urine protein level or urinalysis result?
Infectious	Disease Status (HIV)	Can you tell me if you have HIV, and whether you are currently receiving treatment?	Do you have a diagnosis of Human Immunodeficiency Virus (HIV) infection, and if so, are you on treatment?

Table 4: Example comparison of human-generated and machine-generated patient-facing questions.

cant room for improvement in fully mirroring human phrasing and nuance. This benchmark, therefore, serves its purpose as a critical tool for both validating and iteratively refining generative models for patient-facing communication.

The final benchmark is an assessment corpus designed to evaluate the end-to-end reasoning and classification capabilities of a large language model. This dataset is structured as a collection of tuples: criterion, criterion type, question, user’s answer, comprising 13,344 labeled examples drawn from 1,366 unique clinical trials. Using this benchmark, we quantitatively evaluated the model’s performance against a human-labeled gold standard (Table 6). Across the full evaluation set, the base MedGemma 27B model achieved a three-class accuracy of 74.1% and a binary (non-exclusion vs. exclusion) accuracy of 81.3%, the latter reflecting the model’s tendency to err on the side of caution when a definitive exclusion decision cannot be reliably supported.

The results revealed several systematic patterns. The model performed most reliably on explicit, unambiguous eligibility cases: the INCLUDE class achieved precision 0.83, recall 0.72, and F1 0.77, while EXCLUDE reached precision 0.69, recall 0.82, and F1 0.75. However, performance degraded substantially when faced with ambiguity. The UNKNOWN class represented the key failure point, with precision 0.58, recall 0.64, and F1 0.61 overall—and dropping as low as F1 0.32 on the original dataset where such cases are underrepresented (5.7% of labels), indicating that the model frequently defaulted to a premature binary decision rather than deferring for clarification. This failure mode aligns closely with the qualitative limitations identified in Table 5, including an inability to perform unit conversions for laboratory values (e.g., mmol/L to mg/dL), difficulty interpreting non-clinical patient language (e.g., “within normal range”), and challenges in resolving semantic ambiguity in conditional or tem-

porally qualified statements. This benchmark not only quantifies these predictable weaknesses but also establishes a validated resource for future research, enabling targeted fine-tuning to enhance the clinical assessment capabilities of advanced language models.

5. Discussion

In this paper, we introduce three benchmark datasets designed to advance clinical trial pre-screening research. The paired-criterion benchmark establishes a structured and replicable framework for clustering selection and evaluation. The human-made question benchmark supports the assessment of patient-facing generative AI by providing diverse phrasings of clinical concepts, enabling more nuanced evaluation of question-generation models. The assessment corpus focuses on identifying reasoning challenges such as unit conversion, contextual understanding, and semantic ambiguity. Together, these datasets offer evaluation resources that can inform the development and coordination of future multi-agent chatbots for clinical trial search and recommendation.

Despite these contributions, there are some limitations in our study too. First, our datasets are derived exclusively from five major cancer types. While our data-centric method is designed to be generalizable, the resulting resources are domain-specific. Applying and validating the pipeline in other therapeutic areas will be important to confirm its generalizability. Second, the patient answers in our assessment corpus were synthetically generated by LLM. Although designed to cover diverse response types, they cannot fully capture the complexity, ambiguity, and unpredictability of real-world patient language. Future work include augmentation this corpus with authentic patient dialogues to ensure real-world robustness. These limitations underscore the necessity of expanding data collection

Type of Limitation	Example Question and Answer	Corresponding Criterion	Model Issue / Outcome
Interpretation and Context Understanding	Q: What is your recent blood test that checked how your blood clots, like INR, PT, or aPTT? Were the results normal, and are you currently taking any blood-thinning medications such as warfarin or heparin? A: My blood results are within normal range.	International normalized ratio (INR) and activated partial thromboplastin time (aPTT) or partial thromboplastin time (PTT) > 1.5 × ULN (unless on therapeutic coagulation).	Model fails to infer that “within normal range” implies eligibility and should include all related trials.
Ambiguity in Expression	Q: What is your usual resting heart rate (in beats per minute)? A: My usual resting heart rate is around 72 bpm, though it can increase if I’m stressed or just woken up.	Mean resting heart rate 50–90 bpm (determined from ECG); indication of treatment with hormone therapy and hypofractionated radiotherapy.	Model cannot equate “usual resting heart rate” with “mean resting heart rate,” marking it as “need more information.”
Unit Conversion	Q: What were your most recent fasting cholesterol and triglyceride levels (NCT01301911)? A: Total cholesterol: 185 mg/dL; triglycerides: 110 mg/dL.	Cholesterol 7.75 mmol/L and triglycerides 2.5 × ULN.	Model unable to convert mmol/L to mg/dL (7.75 mmol/L 140 mg/dL), incorrectly marking as “need more information.”
Incomplete Criterion	—	CNS inclusion criteria: [incomplete statement].	Model continues to assess instead of marking as “need more information.”

Table 5: Examples of Model Limitations in Question–Answer–Criterion Assessment

Class	Precision	Recall	F1-score
INCLUDE	0.83	0.72	0.77
EXCLUDE	0.69	0.82	0.75
UNKNOWN	0.58	0.64	0.61

Table 6: Evaluation Performance on Eligibility Assessment

across different diseases and validating our data against real-world interactions.

6. Conclusion

In this paper, we introduced three benchmark datasets that provide the evaluation resources for developing and validating conversational AI in the clinical trial domain. We presented: (1) a paired-criterion dataset for objectively optimizing the clustering of complex eligibility criteria; (2) a human labeled question dataset for the evaluation of patient-

facing language generation; and (3) an assessment corpus that quantifies critical reasoning gaps in current LLMs and enables targeted model improvement, also for further fine-tuning. Ultimately, this work demonstrates that progress in clinical AI is dependent not only on the advancement of the model architecture, but also on the careful curation of high-quality data. These datasets represent a tangible step toward building safer and more reliable patient-centric tools to navigate the complexities of clinical trial recruitment.

7. Acknowledgments

Research reported in this publication was supported by the U.S. National Library of Medicine of the National Institutes of Health under award numbers R01LM014508 and R01LM011934. Yumeng Yang is a predoctoral fellow supported by the Cancer Prevention and Research Institute of Texas (CPRIT).

8. Bibliographical 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 the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.
- Corey W Arnold, Andrea Oh, Shawn Chen, and William Speier. 2016. Evaluating topic model interpretability from a primary care physician perspective. *Computer methods and programs in biomedicine*, 124:67–75.
- Hala T. Borno, Li Zhang, Sylvia Zhang, Celia Kaplan, Alexander Bell, Brian Bakke, Amy Lin, Rahul Aggarwal, and Eric J. Small. 2022. Mobile Clinical Trial Matching Technology in Medical Oncology Clinic: A Pilot Feasibility Study. *JCO Clinical Cancer Informatics*, 6:e2100182.
- Ricardo J.G.B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 160–172. Springer.
- Martin R. Cowie, Juuso I. Blomster, Lesley H. Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, and Mark Leenay, et al. 2017. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1):1–9.
- Betina Ilday, Caitlin Dreisbach, Chunhua Weng, and Rebecca Schnall. 2021. A systematic review on natural language processing systems for eligibility prescreening in clinical research. *Journal of the American Medical Informatics Association*, 29(1):197–206.
- Tian Kang, Shaodian Zhang, Youlan Tang, Gregory W Hruby, Alexander Rusanov, Noémie Elhadad, and Chunhua Weng. 2017. EliIE: An open-source information extraction system for clinical trial eligibility criteria. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Edward S. Kim, Suanna S. Bruinooge, Samantha Roberts, Gwynn Ison, Nancy U. Lin, Lia Gore, Thomas S. Uldrick, Stuart M. Lichtman, Nancy Roach, and Julia A. Beaver, et al. 2017. Broadening Eligibility Criteria to Make Clinical Trials More Representative: American Society of Clinical Oncology and Friends of Cancer Research Joint Research Statement. *Journal of Clinical Oncology*, 35(33):3737–3744.
- Felix Köpcke and Hans-Ulrich Prokosch. 2014. Employing Computers for the Recruitment into Clinical Trials: A Comprehensive Systematic Review. *Journal of Medical Internet Research*, 16(7):e161.
- 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).
- Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, pages 74–81.
- Vivek H Murthy, Harlan M Krumholz, and Cary P Gross. 2004. Participation in Cancer Clinical Trials: Race-, Sex-, and Age-Based Disparities. *JAMA*, 291(22):2720–2726.
- Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolás Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, and Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv*.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, and Elahe Vedadi, et al. 2024. Capabilities of Gemini Models in Medicine. *arXiv preprint arXiv:2404.18416*.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, and Charles Lau, et al. 2025. MedGemma Technical Report. *arXiv*, 2507.05201.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, and Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620.

- Joseph M Unger, Elise Cook, Eric Tai, and Archie Bleyer. 2016. *American Society of Clinical Oncology Educational Book*, volume 36, chapter The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies. ASCO.
- Chunhua Weng, Xiaoying Wu, Zhihui Luo, Mary Regina Boland, Dimitri Theodoratos, and Stephen B Johnson. 2011. EliXR: a system for extracting and encoding eligibility criteria in clinical trials. *Journal of the American Medical Informatics Association*, 18(Supplement 1).
- Yumeng Yang, Ashley Gilliam, Ethan B Ludmir, and Kirk Roberts. 2024a. Exploring the Generalization of Cancer Clinical Trial Eligibility Classifiers across Diseases. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.
- Yumeng Yang, Soumya Jayaraj, Ethan Ludmir, and Kirk Roberts. 2024b. Text Classification of Cancer Clinical Trial Eligibility Criteria. In *Proceedings of the AMIA Annual Symposium*.
- Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, and Chunhua Weng. Criteria2Query: a natural language interface to clinical databases for cohort definition.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the International Conference on Learning Representations*.

HealthTrajectory: Patient Journey Summaries and Visualizations for Patient-Clinician Communication Support

Rohmah Hidayah, Tomohiro Nishiyama, Shoko Wakamiya, Eiji Aramaki

Nara Institute of Science and Technology, Japan

rohmah.hidayah.ri5@naist.ac.jp, {nishiyama.tomohiro.ns5, wakamiya, aramaki}@is.naist.jp

Abstract

In recent years, patient narratives have been used to understand subjective experiences that are not recorded in clinical notes. However, narratives tend to be long and unstructured, requiring summarization. However, text-based summaries often require a lot of clarification from patients and make it difficult for clinicians to review events and changes in symptoms over time. In this study, we expanded the summary output by presenting a visualization of the patient's journey to facilitate communication between patients and medical staff. Referring to the widespread use of LLM for summarization, we compared GPT-4.1 and Gemini-2.5-pro, and used Gemini-3-pro-image-preview for visualization. Data was collected from DIPEX-Japan, then the quality of the summaries was evaluated quantitatively and the visualizations qualitatively. Quantitative evaluation using BLEU and ROUGE metrics showed that Gemini-2.5-pro achieved higher summary scores than GPT-4.1, and Japanese summaries scored higher than English ones. Conversely, English performed better than Japanese in temporal expression extraction using precision, recall, and F1 metrics, and the Gemini-2.5-pro model consistently outperformed GPT-4.1. In qualitative evaluation using the pairwise method, the timetable-based model was far superior with an overall win rate of 0.865 in Japanese and 0.969 in English compared to the baseline.

Keywords: LLMs, Summarization, Visualization, Patient Journey, Timetable

1. Introduction

Sharing information between a patient and a clinician is crucial. The World Health Organization emphasizes the importance of patient and public involvement, positioning patients and families as partners to improve safety, service quality, and patient experience (World Health Organization, 2021). This is especially critical in acute consultations, where clinicians must immediately understand the chronology of symptoms and examinations under tight time constraints.

Patient narratives offer valuable insights into subjective healthcare experiences (Oluoch et al., 2023), yet they are frequently lengthy, fragmented, and unstructured. In particular, chronological events are often expressed inconsistently across languages and contexts, making it difficult for clinicians to synthesize them efficiently (Spitale et al., 2023). In practice, this cognitive load increases the risk of delays and errors in clinical decision-making (Lenz et al., 2025).

Prior work has explored time-based and visualization-oriented approaches. Ledesma et al. (2019) demonstrated that time-based representations assist clinicians in reviewing patient histories and supporting decision-making. Similarly, Torsvik et al. (2025) showed that integrating narrative notes with timelines enables clinicians to quickly access relevant context. Other studies have highlighted the value of patient narratives for understanding healthcare experiences (Ma et al., 2024), time-oriented visualization in clinical con-

texts (Scheer et al., 2022), and information extraction for clinical decision-making support (Elgaar et al., 2025). However, these approaches primarily rely on structured clinical data from healthcare systems, leaving a gap in transforming long, unstructured, patient-generated narratives into outputs tailored for patient-clinician communication.

Our approach focuses on overcoming two primary challenges inherent in this gap:

Extracting dispersed clinical events: Clinically relevant events, ranging from subjective symptom changes to formal medical consultations, are scattered throughout lengthy texts, with their corresponding temporal expressions often being implicit, inconsistent, or context-dependent.

Communication-oriented visualization: While structured tables provide concise data, text-only formats often lack the intuitive clarity required for rapid understanding in clinical settings.

To address these challenges, we introduce **HealthTrajectory**, an LLM-based pipeline designed to convert lengthy, unstructured patient narratives, such as interview transcripts, into an intuitive visual format. The system has two core stages as shown in Figure 1:

Timetable-based summarization: This stage identifies temporal expressions and maps summarized clinical events to these time

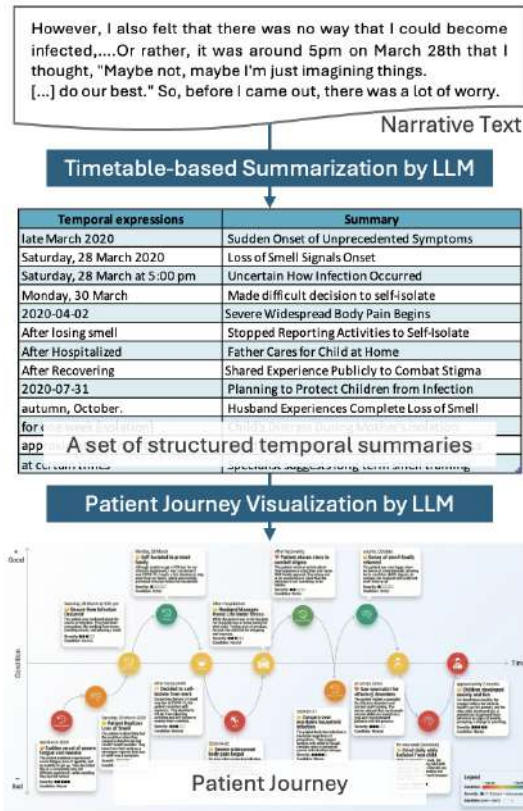


Figure 1: Overview of HealthTrajectory. A patient narrative is processed via Timetable-based Summarization to ensure chronological structure. The resulting summaries are then visualized as a patient journey, transforming raw text into an intuitive trajectory. This intermediate structured step ensures high fidelity, temporal accuracy, and content completeness in the final output.

points to create a structured temporal summaries.

Patient journey visualization: This stage converts the structured summaries into a chronological graphical representation to facilitate effective two-way communication.

The main contributions of this study are as follows:

- We propose **timetable-based summarization**, which identifies temporal expressions and maps summarized clinical events to them, distilling a patient’s fragmented narratives into a single structured table.
- We propose a **patient journey visualization** generated from the structured table, enabling a concise and interpretable presentation that facilitates more effective clinical communication than text alone.

2. Related Work

2.1. Medical Summarization with LLMs

The development of LLM has triggered a surge in medical summarization research because these models can summarize long texts such as visit summaries, emergency room summaries, and discharge summaries. For instance, a study by Williams et al. (2025) evaluating GPT-4 and GPT-3.5 for emergency department visit summaries found that while LLM summaries were reasonably accurate, they remained prone to hallucinations and sometimes omitted relevant clinical information.

Other studies have explored the automation of discharge summaries using open-source models, emphasizing an editable draft approach to ensure clinician supervision and patient safety (Ganzinger et al., 2025). Additionally, Grazhdanski et al. (2025) focused on generating synthetic discharge summaries for training and evaluation purposes to bypass the constraints of real patient data. However, conventional paragraph-based summaries often do not clearly highlight the temporal structure and sequence of events, so the chronology still has to be manually rearranged. Consequently, studies such as Williams et al. (2025) encourage more structured summary formats that facilitate traceability to the original source.

2.2. LLM-based Visualization

Research related to converting summaries into images or visualizations is still limited, especially in the medical field. However, clinical communication often requires quick understanding, and visualization is an important complement to text summaries. In the field of natural language to visualization, studies such as Wu et al. (2024) show that LLMs can help convert sentence instructions into visualizations and empirically evaluate the accuracy of the results.

On the other hand, Ouyang et al. (2025) presents an agent-based workflow approach for NL2Vis and highlights the challenges of cross table reasoning, showing that the ability of LLMs to generate visuals depends on input structure and control mechanisms. At the benchmark level, Chen et al. (2024) also confirms that NL2Vis remains challenging and requires rigorous evaluation in the era of LLMs. However, none of the studies have yet created chronological patient summary visualizations that integrate time and medical event information in a structured manner for ease of communication between patients and clinicians, making the visuals not only appealing but also accountable for clinical discussions.

Entry (Excerpts with annotated temporal expressions)	Gold Summary	Theme
However, I also felt that there was no way that I could become infected, and so, as someone in a reporting position, at the time I was already working from home as much as possible, avoiding rush hour traffic, trying to avoid crowds, and wearing a mask, so I had no idea how I had been infected. Or rather, it was around 5pm on March 28th (Saturday, 28 March at 5:00 pm) that I thought, "Maybe not, maybe I'm just imagining things." [...] So, I called my boss and contacted my husband, and since I was at work for a bit, I went home, and we decided that I had no choice but to contact the public health center and follow their instructions, so I went home by train, wearing a mask and being careful.	I felt there was no way I could get infected. I worked from home as much as possible, avoided rush hour and crowded places, and wore a mask, so I had no idea how I was infected.	Route of infections
After I was hospitalized (After hospitalized) , my husband (who was in close contact with me and is under health observation) stayed at home for two weeks, but we couldn't avoid going shopping, and I couldn't keep my kids confined to the house from morning until night. [...] But even more than that, people whose wives have COVID might be wondering if it's okay for them to be in the park at that moment, or if it's okay to go shopping, but if they don't, they won't be able to eat, and they have children, and so I heard that he was probably thinking about things like that out in the field.	While she was in hospital, her husband was waiting at home and couldn't keep the child locked up all the time, so he chose a time to go shopping and take the child to exercise in the field. I think her husband was also under a lot of pressure.	Young children and coronavirus infection

Table 1: Sample entries from a single patient interview (H01). The table presents English translations of excerpts from the Japanese transcripts, where annotated temporal expressions are highlighted in bold, followed by their normalized labels in parentheses.

3. Dataset

3.1. Materials

Our primary data source comprises patient narratives curated by DIPEX-Japan¹, a national branch of DIPEX International². Adhering to the qualitative research methodology developed at the University of Oxford (Torishima et al., 2020), DIPEX-Japan emphasizes "patient-centered medical care" by documenting the lived experiences of health and illness as a vital social resource.

For this study, we utilized a specialized database of COVID-19 narratives in DIPEX-Japan, comprising interview transcripts that document the lived experiences of 14 individuals: 12 patients (6 men and 6 women, aged 30–70) and 2 patient family members. The interviews were conducted in Japanese from January to July 2021. Note that we excluded two interviews conducted with family members to focus exclusively on first-person patient narratives, resulting in a final dataset of 12 patients.

To ensure thematic consistency, DIPEX-Japan categorized the interview content into 9 themes: "routes of transmission", "children and coronavirus infection", "onset of symptoms", "main symptoms in the acute phase", "anxiety and suffering of in-

Interview	Entries	Words (ja)	Words (en)
H01	13	7,847	5,289
H03	6	3,773	2,086
H04	7	3,395	2,217
H05	10	4,638	2,712
H07	4	1,915	1,208
H08	13	6,570	3,997
H09	7	3,026	1,866
H10	8	4,444	2,653
H11	8	4,588	2,955
H12	8	3,618	2,290
H13	7	3,283	2,284
H14	10	5,789	3,630

Table 2: Statistics of the COVID-19 narrative dataset. The dataset based on 12 interviews (one per patient), with a total of 101 thematic sections (entries) identified across all interviews. While the original database included 14 individuals, two interviews with family members (H02 and H06) were excluded to focus on patient narratives. For each interview, the number of available entries and total text length (word counts) are reported.

fectured patients", "impact on work and the workplace", "long-term symptoms and side effects", reporting infections", and "views on life and the world after COVID-19". Since not every patient discussed all nine themes, we treated each available thematic

¹<https://www.dipex-j.org/>

²<http://www.dipexinternational.org/>

section as an individual entry as shown in Table 1. Consequently, each patient is represented by a varying number of entries as shown in Table 2, each paired with a summary that serves as a gold standard for text summarization.

3.2. Dataset Construction and Annotation

We translated the original Japanese transcripts into English to evaluate the system’s cross-language utility and to ensure consistent performance across different languages. Table 2 provides a statistical overview of the dataset, including the number of patients, segments, and the average text length in both Japanese and English.

Furthermore, to support the evaluation of our patient journey visualizations, one of the authors manually annotated temporal expressions within each segment to form a gold standard timeline. This reference allows for a direct comparison between human-annotated temporal structures and LLM-generated outputs.

4. Method

HealthTrajectory systematically transforms unstructured patient narratives into structured visualizations through a two-stage pipeline. First, we perform timetable-based summarization (Section 4.1), in which an LLM extracts key clinical events and aligns them along chronologically to create structured temporal summaries. Second, these summaries serve as the input for patient journey visualization (Section 4.2), where the LLM renders a graphical representation of the patient’s health trajectory.

4.1. Timetable-based Summarization

The initial stage generates summaries and temporal expressions that maintain chronological coherence, thereby producing a stable and structured foundation for our workflow. The process consists of the following steps:

- **Narrative processing:** The input text (e.g., the patient narrative shown in Table 1) is converted into structured XML format along with a timeline. Our system is implemented in both Japanese and English. Both datasets are processed using the same procedure to ensure a fair bilingual comparison of the LLM’s performance.
- **Summary Generation:** The model generates concise text summaries from the narratives, ensuring a consistent format throughout the entire timeline.

- **Timeline generation:** The LLM generates time-based structured output, specifically identifying events and their corresponding temporal expressions. For temporal expressions, the model extracts exactly as it appears in the narrative if the time is an expression such as “a week later, before testing positive” in order to maintain time consistency and avoid causing confusion.
- **Chronological ordering:** The model arranges events from the earliest to the latest. It prioritizes the inherent temporal order of the narrative without interpolating or inventing non-existent time points.

Finally, the output is checked against the predefined schema and required chronological structure. If inconsistencies are found, such as empty columns, invalid events, repeated content, or incorrect event ordering, the model is prompted to regenerate the output until it matches the required format and temporal sequence.

4.2. Patient Journey Visualization

In this stage, we employ an LLM to generate patient journey visualizations. The structure and components of the prompt used for the visualization are shown in Figure 4. Specifically, the prompt is designed to consider the following points:

- **Chronological order:** The timeline is maintained from the earliest point to the present. This is essential because the present state is often implicit rather than explicitly stated in the narrative.
- **Linear connectivity:** Connections between nodes are restricted to a single, one-way chain without branching to ensure a clear and unambiguous narrative flow.
- **Node composition:** Each node features a card-like display incorporating a temporal label, a concise description, an icon/emoji, and a “severity” indicator to reflect changes in the patient’s condition over time.
- **Readability and layout:** Explicit rules are established to optimize visual clarity, such as the vertical arrangement of nodes (top to bottom), the prevention of overlapping, and the inclusion of legends. These constraints facilitate rapid information processing for clinical decision-making and help patients intuitively understand their disease progression.

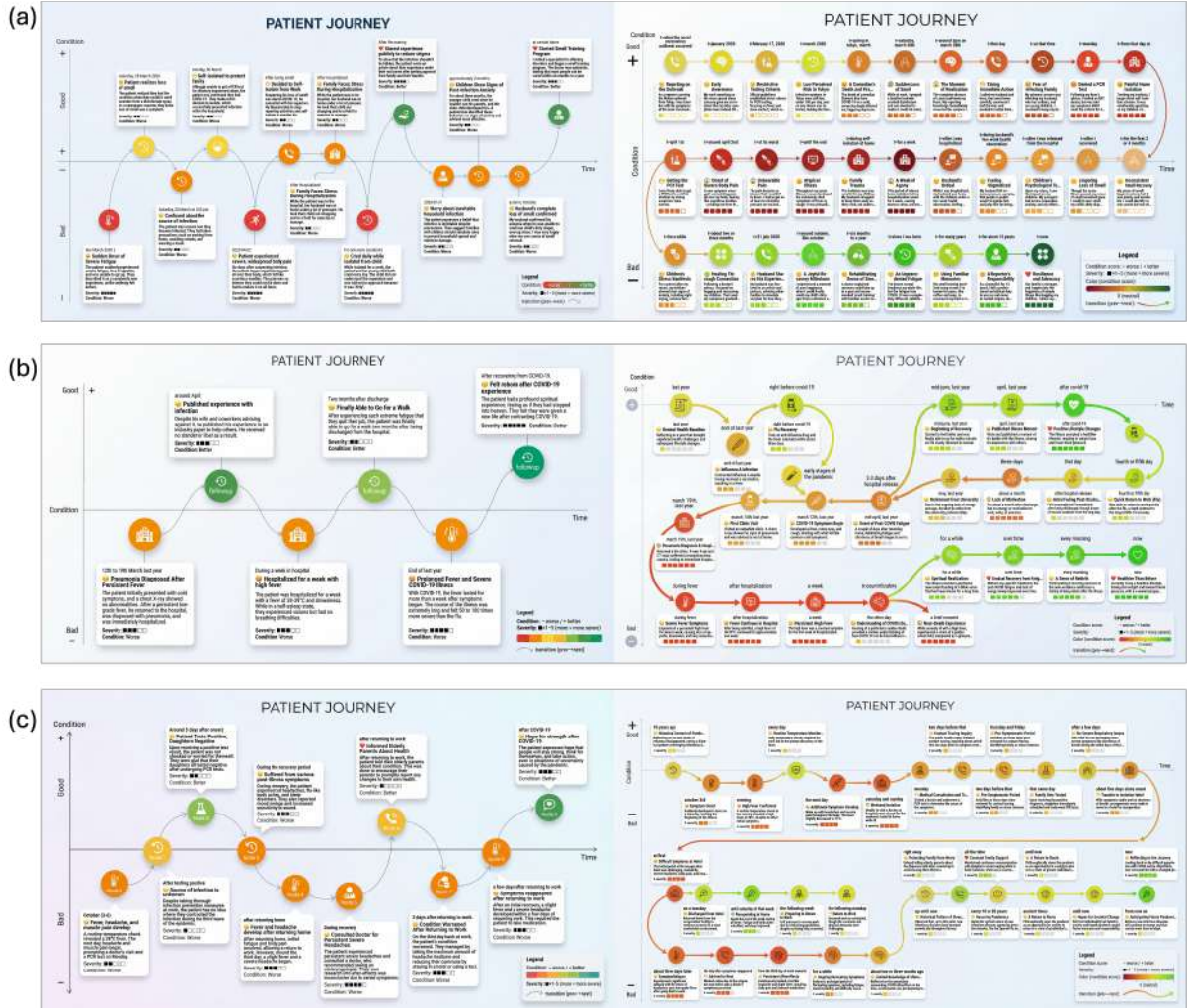


Figure 2: Comparison of patient journey visualizations across three cases. Each pair shows the visualization generated by HealthTrajectory (left) and the baseline (right). This visualization shows the patient’s condition score, where a higher score on the positive y-axis indicates a more positive change in the patient’s condition, meaning the patient is improving. We also use color to facilitate a quicker understanding of the patient’s condition. On the patient card, we also add a number of boxes to the severity scale, indicating a more severe condition if the severity level is high, and a lower number indicates improvement. Transition markers are used to indicate the chronological order of events.

5. Experimental Setup

Our system consists of two steps: experimental setup for timetable-based summarization (Section 5.1) and timetable-based patient journey visualization (Section 5.2).

5.1. Timetable-based Summarization Setup

5.1.1. Models and Prompts

We chose the Gemini-2.5-pro and GPT-4.1 models with a zero-shot setting. We selected GPT-4.1 because medical evaluations based on rubrics indicate that GPT-4.1 is a high-performing model, especially for clinical text comprehension tasks that

require accuracy and can perform consistent temporal extraction (not fabricating time and maintaining the sequence of events) (Arora et al., 2025). We chose Gemini-2.5-Pro because official technical reports demonstrate its advanced capabilities in reasoning, multimodality, and long context (Comanici et al., 2025). Both are highly suitable for our workflow, which processes long patient narrative texts into timetable-based summarization.

We used a prompt that instructs the model to divide a single interview transcript from one patient into several segments in chronological order, considering the aspects found in the patient’s narrative data as shown in Figure 3. For each segment, the model is instructed to extract a timeline with one to two TIMEX3-style temporal expressions, with con-

System prompt: “You are a time-label sorter, specialized in ordering timeline labels into a strict chronological sequence (oldest→newest).”

Instruction: You receive one interview represented as multiple timeline entries, where each entry contains a timeline label and an optional aspect field. Sort the entries strictly from **oldest to newest**, ignoring the original table order and narrative/story order. If a timeline label is vague, you may use the aspect field only as supporting evidence, but you must not hallucinate or infer dates/events that are not stated in the input. Expressions such as “after N years ...” are not concrete dates and should be placed after entries with explicit dates, while “now/recently” must be placed as the newest.

Output: Return a **JSON** object that provides the **chronological ordering** of all entry indices from oldest to newest.

System prompt: “You are a medical summarizer, specialized in producing faithful point-level summaries for patient-journey tracking.”

Instruction: You receive a time label (which must be used **as-is**), an optional aspect field (supporting cue only), and the corresponding content (a text segment or an existing summary, when available). Generate **exactly one** timetable point for each timeline entry. Do not add any facts that are not present in the input (**no hallucination**). Write a short medical summary consisting of a concise `heading` and `detail`. The `heading` must be short (≤ 7 words) and complete (no ellipsis). The `detail` must be concise and complete, written as 1–2 full sentences (no ellipsis, no truncation). Select `emoji` from a predefined emoji pool, and `icon_hint` from a predefined icon inventory. Assign a `severity` level on a 1–5 scale, where 1 indicates the **worst** condition and 5 indicates the **best** condition. If the generated text appears cut off or contains ellipses (“.../...”), rewrite it to be shorter but still complete, while preserving the original meaning.

Output: Return a **JSON** object that contains the medical summary and its **severity** level (1–5), along with auxiliary fields used for visual rendering.

Figure 3: Prompts for timetable-based summarization.

System prompt: “You are a medical infographic designer, specialized in rendering a constrained patient-journey timeline as a clean, readable 16:9 vector infographic.”

Instruction: You receive a chronologically ordered list of patient-journey points and must produce a professional **16:9** infographic in a clean **vector** style with sharp, readable text (no blur). The infographic must contain exactly n nodes (colored circles) and exactly $n-1$ arrows, forming a single chain $0 \rightarrow 1 \rightarrow 2 \rightarrow \dots \rightarrow n-1$ with no branching and no extra icons/circles. Draw thin axes: a thin x-axis across the middle labelled *Time* near the right end and a thin y-axis on the left labelled *Condition*. Reserve a clean left margin lane ($\sim 10\%$ of the canvas width) exclusively for the y-axis label and polarity markers; no nodes, arrows, cards, legend, or other objects may overlap this lane. Place nodes above/below the x-axis based on the qualitative condition label (e.g., Better/Worse/Neutral), and never print any hidden numeric variables used for layout. Use the provided `severity` (1–5) to encode visual emphasis (e.g., color intensity or a 5-block bar), where 1 represents the **worst** condition and 5 represents the **best** condition. Every node must have exactly one attached card connected by a thin leader line, and cards must not overlap each other, nodes, arrows, axes, the reserved lane, or the legend. Each card must include the time label (as provided), bold `emoji` + `heading`, the `detail` text (1–2 complete sentences), the `severity` indicator (five blocks), and a qualitative label *Better/Worse/Neutral*. Include a compact legend at the bottom-right showing the severity key (1–5 blocks) and one sample thin curved arrow labelled *transition* (*prev*→*next*); do not show hex codes and do not show any hidden layout scores.

Output: Return one **16:9** infographic image satisfying all constraints above.

Figure 4: Prompt for patient journey visualization.

tent length adjusted to the medical context of each segment. Additionally, the model is asked to summarize the narrative into one to three sentences that align with the medical context of each segment without adding new information.

5.1.2. Evaluation Metrics

Automatic evaluation consists of two components: summary evaluation and timeline evaluation.

Summary Evaluation: We use ROUGE-1/2/L to capture diverse aspects of summary quality. ROUGE-1 (unigram) evaluates how much information from the reference is captured, while ROUGE-2

(bigram) evaluates more strictly and ensures that the model not only captures information but also retains phrases. ROUGE-L (longest common subsequence) assesses similarity in terms of event chronology more flexibly (Zhang et al., 2025). Additionally, we use BLEU to provide a complementary perspective on n-gram overlap (Chu et al., 2024).

Timeline Evaluation: For the timeline generation task, we use precision, recall, and F1-score. These metrics are essential for ensuring that the model accurately extracts temporal expressions while minimizing the inclusion of incorrect time points. This approach is often used for evaluating

temporal information extraction, which are often referred to as TIMEX3 (Yao et al., 2025).

5.2. Patient Journey Visualization Setup

5.2.1. Models and Prompts

We used the Gemini-3-pro-image-preview model³⁴. This model was launched on November 20, 2025, as the final iteration of the Nano Banana model for image generation. This model is designed to follow complex instructions with more accurate image results and is specifically tailored for the production of high-quality visual assets, which is why we chose this model over others.

We use a prompt that instructs the model to visualize the patient's journey with a timeline infographic design that includes visual elements such as icons, nodes, and chronological order, as shown in Figure 4. The prompt also instructs the model to create a time axis (x-axis) and a condition axis (y-axis), and each event has a time label, an icon/emoji as a marker, a brief description, and a severity level indicator (1–5) shown using different colors. To keep the visualization readable and avoid overlapping when many events are present, the layout is allowed to extend downward. Overall, the instructions aim to keep all elements structured and clearly presented.

5.2.2. Evaluation Metrics

To quantify the effectiveness of the generated visualizations, we conducted a human evaluation based on pairwise comparisons over the entire dataset. This method provides a more reliable measure of human preference than independent scoring, as it directly compares the outputs of two systems given the same input narrative.

We evaluated the visualizations using three criteria, each consisting of two evaluation aspects summarized from our questionnaire:

- **Correctness:** Assesses *clinical relevance* (whether the visualization accurately reflects the medical context) and *hallucination avoidance* (whether the visualization avoids adding non-existent information).
- **Coverage:** Evaluates *completeness* (the extent to which information from the table is captured) and *faithfulness* (the degree to which the visualization adheres to the table).
- **Visual design:** Measures *readability* (ease of understanding the content) and *layout quality*

³⁴www.ai.google.dev/gemini-api/docs/image-generation

⁴www.ai.google.dev/gemini-api/docs/changelog

(effectiveness of the design for following the timeline).

To determine the relative performance, we calculated the winning proportion for the proposed method. Let W_{base} denote the number of votes for the baseline, and W_{ours} denote the votes for the proposed method. The winning proportion is defined as:

$$\text{WinProp} = \frac{W_{ours}}{W_{base} + W_{ours}}, \quad (1)$$

where $\text{WinProp} > 0.5$ indicates a preference for the proposed method.

The evaluation was conducted by four Japanese evaluators, including three NLP researchers and one medical professional, to ensure both technical and clinical validity. For each patient case, they were presented with the gold standard timeline table (Section 3.2) and asked to determine which visualization better represented the structured information.

6. Results and Discussion

6.1. Timetable-based Summarization

In timetable-based summarization, we evaluated two viewpoints: the results of the Gemini-2.5-Pro model and the GPT-4.1 model on the timeline, as well as the results of the Gemini-2.5-Pro model and the GPT-4.1 model on summarization.

Summarization Results: Table 3 shows the summarization evaluation results. In the table, the Gemini-2.5-pro model achieved higher scores in both Japanese and English compared to the GPT-4.1 model in the narrative summarization task. Specifically, the Gemini-2.5-pro model scored the highest in all measurements for Japanese, with a score of 0.506 on ROUGE-1, followed by a score of 0.230 on ROUGE-2, then 0.316 on ROUGE-L, and 0.474 on BLEU. This indicates that the Gemini-2.5-pro model is more stable in maintaining core content and word choices that are closer to the reference.

Timeline Results: Table 3 shows the performance of the timeline generation. The results in English are higher compared to Japanese, which is in contrast to the summary evaluation that shows higher scores for Japanese compared to English. The scores for precision, recall, and F1 are consistently highest for the Gemini-2.5-pro model, with a score of 0.251 for precision and 0.242 for recall and 0.246 for F1 in English time extraction. From these results, the Gemini-2.5-pro model appears more conservative in extracting times that closely match the reference compared to the GPT-4.1 model.

Lang	Model	Summary Evaluation				Timeline Evaluation		
		ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	BLEU↑	P↑	R↑	F1↑
JA	Gemini-2.5-pro	0.506	0.230	0.316	0.474	0.157	0.157	0.157
JA	GPT-4.1	0.472	0.204	0.294	0.433	0.179	0.159	0.168
EN	Gemini-2.5-pro	0.276	0.077	0.186	0.253	0.251	0.242	0.246
EN	GPT-4.1	0.256	0.065	0.176	0.233	0.250	0.231	0.240

Table 3: Performance of timetable-based summarization in Japanese (ja) and English (en) text.

Criterion	Aspect	Japanese			English		
		W_{base}	W_{ours}	WinProp	W_{base}	W_{ours}	WinProp
Correctness	Clinical relevance	8	40	0.833	1	47	0.979
	Hallucination avoidance	3	45	0.938	0	48	1.000
Coverage	Completeness	1	47	0.979	0	48	1.000
	Faithfulness	3	45	0.938	0	48	1.000
Visual design	Readability	13	35	0.729	5	43	0.896
	Layout quality	11	37	0.771	3	45	0.938
Overall		39	249	0.865	9	279	0.969

Table 4: Results of Patient Journey Visualization using the pairwise method. W_{base} and W_{ours} show the vote counts for the baseline and the proposed method by aspect and language, respectively; WinProp is the winning proportion of our method.

6.2. Patient Journey Visualization

As shown in Table 4, the evaluators consistently preferred the proposed method over the baseline across all evaluation aspects for both Japanese and English. The proposed method achieved substantially higher preference scores, with 0.865 for Japanese and 0.969 for English, indicating a clear advantage over the baseline. These results suggest that structuring patient narratives into timetable-based visualizations improves the perceived quality of the output in terms of clinical relevance, completeness, faithfulness, and hallucination reduction.

in the English setting, the proposed method achieved perfect scores (1.000) for hallucination avoidance, completeness, and faithfulness. In contrast, improvements in visual aspects such as readability, layout quality, and clinical relevance were more moderate, particularly in the Japanese setting. This discrepancy may indicate that while the proposed method effectively constrains content generation, visual presentation quality is influenced by additional factors, such as language-specific phrasing and formatting preferences.

Overall, the proposed method (left panel of Figure 2) appears to be well-suited for bilingual Patient Journey visualization, supporting clearer and more structured explanations of patient trajectories. The results based on Gemini-2.5-pro further suggest that high-quality upstream generation is critical for producing coherent and interpretable visualizations, particularly in maintaining alignment between clinical

events and temporal information.

7. Conclusion

We proposed a bilingual timetable-based Patient Journey system that generates structured visualizations from patient narratives to support clinician-patient communication. The evaluation results demonstrate that the proposed system is robust for summary generation, with the Japanese version generally outperforming the English version on both the Gemini-2.5-pro and GPT-4.1 models. In contrast, for temporal expression extraction, the English version achieved higher accuracy, and Gemini-2.5-pro consistently outperformed GPT-4.1. Furthermore, in the human evaluation comparing the baseline and the proposed method, the evaluators almost always chose the proposed method. This preference was primarily attributed to its higher clinical relevance, fewer hallucinations, greater completeness, and improved readability, making it more suitable for supporting communication between patients and clinicians across languages.

In future work, we plan to implement **HealthTrajectory** for direct use in clinician-patient communication. This system will provide an interface that enables patients to upload information about their symptoms prior to pain onset. The system will generate a visual summary and timeline of the patient's journey. The system will also include features that support both patients and clinicians, allowing the outputs to be reviewed prior to use in clinical discussions.

8. Limitations

There are several limitations to this study. First, the experiments were conducted on a small dataset of COVID-19 patient narratives, requiring validation on a larger and more diverse dataset, including other diseases and narrative types. Second, the timeline annotation was performed by only one annotator, leading to a lack of agreement between annotators and also introducing the potential risk of subjective bias in the gold standard annotation. Third, for the generated patient journey images, each case was generated up to four times as a practical measure to reduce inconsistencies; however, some results still showed inconsistencies regarding severity. Furthermore, the current system is still in its early research phase, and these findings are intended to inform further development towards more practical use in the future.

9. Acknowledgements

This work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP), Project JPJ012425, and the Japan Society for the Promotion of Science (JSPS) Research Start-up Grant JP25K24412. We also sincerely thank the JICA Research Scholarship for their financial support and our research group members for their useful discussions and comments.

10. Bibliographical References

- Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.
- Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. 2024. Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics*.
- An Chu, Thong Huynh, Long Nguyen, and Dinh Dien. 2024. A comparative study of chart summarization. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 971–981.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- 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.
- Matthias Ganzinger, Nicola Kunz, Pascal Fuchs, Cornelia K Lyu, Martin Loos, Martin Dugas, and Thomas M Pausch. 2025. Automated generation of discharge summaries: leveraging large language models with clinical data. *Scientific Reports*, 15(1):16466.
- Georgi Grazhdanski, Vasil Vasilev, Sylvia Vassileva, Dimitar Taskov, Izabel Antova, Ivan Koychev, and Svetla Boytcheva. 2025. Synthmedic: utilizing large language models for synthetic discharge summary generation, correction and validation. *Journal of Biomedical Informatics*, page 104906.
- Andres Ledesma, Niranjan Bidargaddi, Jörg Strobel, Geoffrey Schrader, Hannu Nieminen, Ilkka Korhonen, and Miikka Hermes. 2019. Health timeline: an insight-based study of a timeline visualization of clinical data. *BMC medical informatics and decision making*, 19(1):170.
- Janek Lenz, Isabel Richter, and Sven Meister. 2025. Automated filtering and visualization of patient-centered data from electronic health records in emergency care: A scoping review. *Journal of Multidisciplinary Healthcare*, pages 6503–6517.
- Hua Ma, Xiaoru Yuan, Xu Sun, Glyn Lawson, and Qingfeng Wang. 2024. Seeing your stories: visualization for narrative medicine. *Health Data Science*, 4:0103.
- Dorothy Oluoch, Sassy Molyneux, Mwanamvua Boga, Justinah Maluni, Florence Murila, Caroline Jones, Sue Ziebland, Mike English, and Lisa Hinton. 2023. Not just surveys and indicators: narratives capture what really matters for health system strengthening. *The Lancet Global Health*, 11(9):e1459–e1463.
- Geliang Ouyang, Jingyao Chen, Zhihe Nie, Yi Gui, Yao Wan, Hongyu Zhang, and Dongping Chen. 2025. nvagent: Automated data visualization from natural language via collaborative agent workflow. *arXiv preprint arXiv:2502.05036*.
- Jan Scheer, Alisa Volkert, Nicolas Brich, Lina Weinert, Nandhini Santhanam, Michael Krone, Thomas Ganslandt, Martin Boeker, and Till

- Nagel. 2022. Visualization techniques of time-oriented data for the comparison of single patients with multiple patients or cohorts: Scoping review. *Journal of medical Internet research*, 24(10):e38041.
- Giovanni Spitale, Andrea Glässer, Mirriam Tyebally-Fang, Corine Mouton Dorey, and Nikola Biller-Andorno. 2023. Patient narratives—a still undervalued resource for healthcare improvement. *Swiss Medical Weekly*, 153:40022.
- Masako Torishima, Michiko Urao, Takeo Nakayama, and Shinji Kosugi. 2020. Negative recollections regarding doctor–patient interactions among men receiving a prostate cancer diagnosis: a qualitative study of patient experiences in japan. *BMJ open*, 10(1):e032251.
- Torbjørn Torsvik, Andreas Brun, Bjørn Ståle Benjaminsen, and Aslak Steinsbekk. 2025. Labvis: usability testing of a prototype tool for integrating timeline graphs and clinical notes. *BMC Medical Informatics and Decision Making*, 25(1):337.
- Christopher YK Williams, Jaskaran Bains, Tianyu Tang, Kishan Patel, Alexa N Lucas, Fiona Chen, Brenda Y Miao, Atul J Butte, and Aaron E Kornblith. 2025. Evaluating large language models for drafting emergency department encounter summaries. *PLOS digital health*, 4(6):e0000899.
- World Health Organization. 2021. *Global patient safety action plan 2021-2030: towards eliminating avoidable harm in health care*. World Health Organization.
- Yang Wu, Yao Wan, Hongyu Zhang, Yulei Sui, Wucan Wei, Wei Zhao, Guandong Xu, and Hai Jin. 2024. Automated data visualization from natural language via large language models: An exploratory study. *Proceedings of the ACM on Management of Data*, 2(3):1–28.
- Jiarui Yao, Eli Goldner, Harry Hochheiser, Sean Finan, John Levander, David Harris, Piet C de Groen, Elizabeth Buchbinder, Danielle Bitterman, Jeremy L Warner, et al. 2025. Systemic anticancer therapy timelines extraction from electronic medical records text: algorithm development and validation. *JMIR Bioinformatics and Biotechnology*, 6(1):e67801.
- Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2025. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41.

Medical-FLAVORS-AECC: Spanish Oncological Metaphors Dataset

Lucía Pitarch¹, Jordi Bernad¹, Sergio-Luis Ojeda-Trueba²
Alec Sánchez-Montero³, Maxim Ionov¹, Emma Anglés-Herrero¹
Ángel Óscar Corona Beomont⁴, Gemma Bel-Enguix²

Universidad de Zaragoza, Zaragoza (Spain)¹

Universidad Nacional Autónoma de México, Ciudad de México (Mexico)²

Universidad Pompeu Fabra, Barcelona (Spain)³

Centro Universitario de la Defensa de Zaragoza, Zaragoza (Spain)⁴

{lpitarch,jbernad,mionov,eangles,acorona}@unizar.es

sergioluis0906@gmail.com, alecm.montero@gmail.com, gbele@ingen.unam.mx

Abstract

Metaphors play a central role in cancer narratives, helping patients and practitioners articulate complex experiences and technical concepts. While cancer metaphors in English have been extensively studied, Spanish remains underexplored in this regard, despite its global importance and rich cultural variation. This paper presents a new dataset of Spanish cancer metaphors designed to address these gaps. The resource comprises over 80K annotated words drawn from diverse forum posts, with detailed documentation of lexical units, contextual versus basic meanings, and inter-annotator agreements. To construct the dataset, we adapted the Metaphor Identification Procedure (MIP) for Spanish medical discourse, proposing methodological refinements to challenges such as defining lexical units or domain-specific Basic Meaning labels.

Keywords: Metaphor annotation, Spanish Language Resources, Oncological Metaphors, Cancer Narratives

1. Introduction

Metaphors are cognitive and linguistic devices that enable speakers to explain complex or abstract concepts through more concrete or familiar terms (Lakoff and Johnson, 1980). In highly technical and emotionally sensitive contexts, such as cancer communication, metaphors play a central role in shaping how patients, families, and professionals conceptualize the illness (Pinheiro et al., 2017; Landau et al., 2019; Gustafsson et al., 2019; Fadul et al., 2009). Building on the premise that metaphors are not inherently positive or negative, Semino and Demjen (2017) developed the Metaphor Menu to help patients and practitioners select more suitable language based on individual experiences and preferences, as understanding their use, potential impact, and the patient preferences is key to effective communication.

This paper presents a manually annotated subset from the AECC (Asociación Española Contra el Cancer) forums, a central place for communication among Spanish-speaking cancer patients and their family members. This dataset will be used to train and evaluate a model for the automatic extraction of metaphors from the rest of the corpus. The annotated subset comprises 72 posts (700 words each on average, including comments).

Following a Spanish adaptation of the MIPVU (Metaphor Identification Procedure Vrije Universiteit) (Steen et al., 2010) adjusted for the medical

domain, the subset was annotated by five independent annotators. From over 83,000 words, approximately 2,035 phrases (ca. 2.45%) were labeled as metaphoric vehicles related to the cancer experience by at least one of the annotators. With an overall inter-annotator agreement (IAA) of 0.49 (F1 score), the resulting dataset is available upon request and can be used to fine-tune language models for metaphor identification (see Ge et al. (2023) for possible metaphor identification methods).

The main contributions of this paper are as follows:

- We present Medical-FLAVORS-AECC dataset, a large-scale Spanish-language resource on metaphors used by patients and their families to describe the cancer experience.
- We describe the annotation process and inter-annotator agreement, providing the first resource for the Spanish medical domain that explicitly documents the distinction between a lexical unit's basic and contextual meaning as established by the MIPVU guidelines.
- We provide a curated vocabulary of over 800 distinct metaphorical vehicles (grouped into 81 source and 46 target domains), useful for both qualitative linguistic analysis and NLP applications.

All annotation guidelines, metadata, and the curated vocabulary are released in a GitHub repos-

itory¹ to ensure reproducibility and support future corpus-based and computational studies on metaphor in the medical domain.

This paper begins by reviewing relevant precedents in Section 2. We then describe the data extraction from the AECC forums and present general statistics of the corpus in Section 3, followed by a detailed account of the annotation and curation process in Sections 4 and 5. Finally, in Section 6 we present an analysis of the resulting dataset, highlighting the insights it provides into the cancer experience as communicated by real people.

2. Related Work

Metaphors are not only pervasive in medical language (Casarett et al., 2010; Semino et al., 2017) but also highly relevant. In medical settings, metaphors translate specialized terminology into lay terms (Pinheiro et al., 2017), facilitate decision-making (Landau et al., 2019), support coping (Gustafsson et al., 2019), and address sensitive topics (Fadul et al., 2009).

Research on metaphors has largely focused on English, with limited attention to Spanish, despite its global reach. Authors such as Liu et al. (2024); Pitarch et al. (2024); Magaña (2020) highlight the need to adapt metaphor studies to different languages and varieties. While some resources already exist for Spanish, they are scarce and limited. For instance, the MetaNet project (Dodge et al., 2015) includes only a small and heterogeneous set of Spanish examples, while the first version of Medical-FLAVORS (Pitarch et al., 2024), a dataset of Spanish cancer metaphors, is biased toward Chilean Reddit data, with uneven cancer types and slang-heavy language. Both resources lack transparent annotation procedures, a crucial issue given the subjectivity of metaphor analysis (Sánchez-Montero et al., 2025a,b).

Our work addresses this gap by introducing Medical-FLAVORS-AECC, a new module for the Medical-FLAVORS dataset. For its annotation, we adapt the Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007) to Spanish medical discourse. This method, along with its extension MIPVU (Steen et al., 2010), is the most established framework for metaphor annotation and involves four steps: (1) segmenting the text into lexical units, (2) determining each unit’s basic meaning (the more concrete, physical, or older sense), (3) identifying its contextual meaning, and (4) labeling the unit as metaphorical if a contrast exists between the two meanings.

While extensively used, the MIP poses significant challenges, particularly in defining text segmenta-

tion into relevant lexical units and selecting the most basic meaning of a word. These core decisions are often controversial (De Backer et al. (2023); Maudslay and Teufel (2022)), lack objective metrics, and are rarely documented in new corpora annotations. Furthermore, both MIP and MIPVU were initially developed for English, lacking fine-grained adaptations to Spanish. To our knowledge, the dataset presented in this paper is the first of its kind to address these issues for Spanish by providing explicit guidelines, annotations, and metrics in terms of lexical unit segmentation and meaning distinction.

3. Data Extraction

The data were extracted from the web forums of the Asociación Española Contra el Cáncer (AECC), founded in 1953 to provide assistance and support to cancer patients and their families. Although based in Spain, the forums are accessible worldwide, enabling Spanish-speaking users from diverse regions to share their experiences. It is organized into multiple topics, ranging from specific cancer types to prevention, volunteering, and events. While anyone may read the forums, registration is required to participate. The most active and general thread, “Testimonios de cáncer” (cancer stories), has received contributions continuously since 2010. Given the association’s prominence and its broad coverage of patients, relatives, and experiences over time, the AECC forums were chosen as a representative and ethically appropriate source of cancer discourse in Peninsular Spanish.

We selected posts from three topics: “Testimonios” (general thread), “Cáncer de mama” (breast cancer), and “Cáncer de próstata” (prostate cancer). Their high incidence, together with their cultural and emotional salience (World Health Organization, 2025; Sociedad Española de Oncología Médica, 2025a,b) in patient discourse, make these two types particularly relevant for metaphor analysis in Spanish. The “Testimonios” thread, on the other hand, provides a more general overview of first-hand cancer experiences. Each post in the forums consists of a title, an initial message from the main author, and a thread of subsequent comments from other community members.

4. Annotation

The annotation process was carried out in two phases: a pilot testing phase for guideline refinement, followed by the main annotation phase. In the pilot testing, all annotators applied a preliminary draft of the guidelines to the first five texts. Questions and disagreements were discussed in a subsequent meeting and integrated into the revised guidelines, with special attention to the delimitation

¹<https://github.com/LuciaPitarch/MedicalFLAVORS-AECC/>

of lexical units. To ensure consensus, a sixth text was co-annotated and reviewed by the entire team. For the main annotation, the remaining posts were distributed among annotators under the following conditions: (a) each post was annotated by two annotators, (b) the workload was balanced, and (c) all possible annotator pairs were assigned common posts. The annotation took place between May and October 2025, following the principles of the Belmont Report ([National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979](#)). A summary of the annotators' profiles can be found in Table 3.

4.1. Annotation Platform and Guidelines

This section provides a summary of our annotation guidelines. The entire annotation was conducted on the INCEpTION platform, as illustrated in Figure 1, where we can observe how each annotated post follows a typical forum structure: an initial message from the main author followed by a series of comments from other community members. Figure 2 summarizes the annotation pipeline we applied.

4.1.1. Definition of Metaphor

Following the principles of MIP ([Pragglejaz Group, 2007](#)) and its extension MIPVU ([Steen et al., 2010](#)), we define a metaphor as a linguistic expression in which a more concrete, physical, or perceptually imaginable concept is used to refer to a more abstract domain of experience. For instance, describing CANCER AS A BATTLE constitutes a metaphor, as it draws on the tangible and culturally salient domain of warfare to frame the intangible and multifaceted experience of illness.

4.1.2. Lexical Unit Segmentation

Given the complexity of metaphorical expressions in our dataset, a flexible approach to lexical unit segmentation was required. Metaphorical meaning often arises not from isolated words, but from the interaction of multiple elements within a phrase or clause, or even from a text as a unit. It can sometimes only be comprehended by comparison with very similar clauses where different words interact.

For example, in constructions such as *voy paso a paso con el tratamiento* ("I go step by step with the treatment") or *lucho a diario contra el cáncer* ("I fight daily against cancer"), the metaphor is distributed across several words. Similarly, in expressions like *recibir tratamiento* ("to receive treatment") or *pasar por tratamiento* ("to go through treatment"), the metaphorical meaning, where treatment is conceptualized as an object or a milestone in a journey, emerges from the semantic interaction among the

elements involved. In such cases, understanding how the person envisions the experience of being ill with cancer or how they conceptualize treatment only becomes possible by analyzing the full clause.

To account for this, we adopt Lynne Cameron's approach to metaphor annotation ([Cameron, 2007](#)), particularly the notion of *metaphorical vehicles*: clusters of words that collectively convey a metaphorical mapping. We annotate all lexical elements central to the metaphor, including verbs, nouns, adjectives, and multiword expressions.

Importantly, only metaphorical expressions that pertain to the experience of the illness (e.g., physical, emotional, or cognitive aspects of cancer) are annotated, which means our focus is on cancer metaphors. When the metaphor is confined to a single word, it is annotated in the `Metaphors` layer, as depicted in Figure 2.

4.1.3. Annotating Basic and Contextual Meaning

Although MIP and MIPVU highlight the centrality of the *basic meaning* (BM) and *contextual meaning* (CM) in metaphor identification, few concrete examples are provided in their documentation. Besides, BM and CM are not explicitly annotated in most publicly available datasets. To enhance transparency and reproducibility in our annotations, we chose to explicitly annotate the BM and CM for each metaphorical unit. To do so, we embedded the Spanish version of the Open Multilingual Wordnet ([Vossen et al., 2016](#)) into the INCEpTION platform, creating an integrated tool for annotation specifically adapted to our task.

The annotators were instructed to use the BM annotation fields as follows:

- Enter the basic meaning in the `Basic-Meaning` field.
- When available, retrieve definitions from WordNet that correspond to the most concrete or prototypical sense of the lexical unit.
- If no appropriate WordNet entry exists, consult the *Diccionario de la lengua española* (RAE) ([Real Academia Española, 2023](#)) or other authoritative sources such as ASALE ([Asociación de Academias de la Lengua Española, 2023](#)).
- If neither resource provides a suitable definition, provide a manually formulated paraphrase based on your interpretation of the basic meaning.
- Use the `Contextual-Meaning` field to define the meaning of the lexical unit in the given utterance. If a figurative sense is listed in the

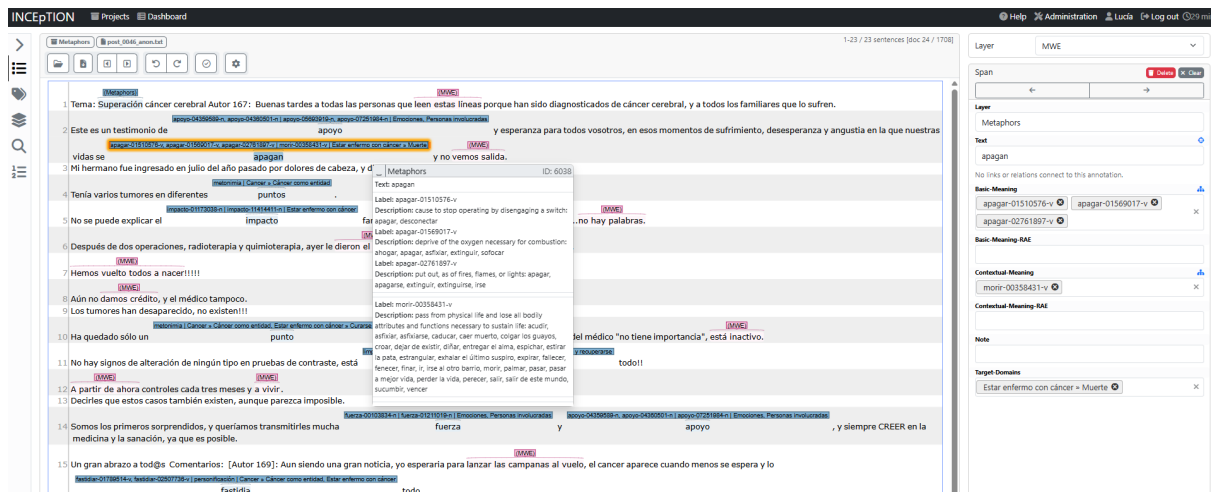


Figure 1: Annotation in INCEPTION, with the additional lexical resources at right

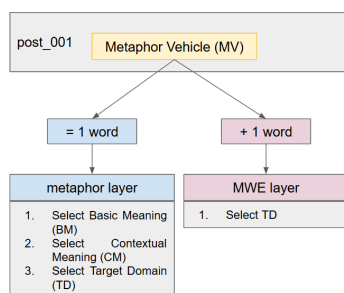


Figure 2: Schematic annotation pipeline

lexical resources, select it accordingly. When no direct match exists, but related entries reflect the intended sense, they may be adapted and cited.

These annotations can facilitate further semantic analysis by leveraging hypernymy/hyponymy relations and identifying recurring conceptual domains.

4.1.4. Identifying Cancer-Related Themes

We tagged each identified metaphor with its corresponding target domain(s) to classify the aspect of the cancer experience it addresses. Common domains include bodily experience, treatment, emotional impact, and social relationships.

Annotators selected the most appropriate theme(s) from a predefined list in the *Target-Domain* field. Multiple themes could be tagged if a metaphor spanned more than one dimension of experience. The topics were extracted from a relevant scoping review on cancer metaphors (Liu et al., 2024) and further adapted to gain a finer understanding of the experience through discussions with healthcare professionals and conversations with patients. The full list of topics is available on

GitHub.

4.1.5. Notes and Issue Tracking

Annotators are encouraged to use the *Note* field (in both the *Metaphors* and *MWE* layers) to document uncertainties, interpretation challenges, or contextual ambiguities encountered during the annotation process. In the event of repeated content, textual errors, or systemic inconsistencies, annotators had to report them in the shared *Incidencias* log for follow-up and dataset revision.

4.2. Agreement Metrics

To report our inter-annotator agreement (IAA) we provide F1 score, following Boguslav and Cohen (2017), as it has been proven to be a more transparent and efficient metric for tasks that involve phrase extraction. Most widely used metrics to report IAA, such as Cohen's Kappa or Krippendorff's Alpha, will not work well on tasks where many spans are possible and the large number of possible negative spans can distort agreement scores.

Using F1 score implies that, for each post, one annotator's labels were treated as the gold standard and the other's as the prediction. An agreement, or a true positive, was defined by a partial overlap, meaning at least one word was shared between the spans marked by the two annotators in any layer (metaphor or MWE). For instance, if Annotator A marked "fight against cancer" and Annotator B marked only "fight", the overlap was considered as an instance of agreement. The overall F1 score was 0.49. A detailed breakdown of the results is presented in Tables 1 and 2.

Annotator	Mean F1
A1	0.60
A2	0.52
A3	0.48
A4	0.53
A5	0.49

Table 1: Mean IAA scores by annotator

L ₁	L ₂	F1	L ₁	L ₂	F1
A2	A1	0.64	A3	A1	0.62
A1	A4	0.60	A5	A1	0.55
A2	A3	0.54	A2	A4	0.47
A3	A5	0.47	A2	A5	0.46
A5	A4	0.46	A3	A4	0.45

Table 2: IAA scores by annotator pair (L₁,L₂)

5. Curation

The curation process of the annotated dataset focused on two main challenges: (1) clustering metaphorical vehicles (including whether to analyze them compositionally or holistically), and (2) refining the final annotations to determine which specific words would be labeled as metaphorical.

5.1. Curation of Metaphorical Vehicles

Metaphorical vehicles pose particular challenges for annotation due to their morphological variability, syntactic discontinuity, and position on a continuum from compositional to fixed expressions. In this work, we distinguish two main categories of multi-word metaphorical vehicles: (a) **Multi-Word Expressions (MWEs)**, understood as conventionalized lexical combinations with fixed or semi-fixed structures (Corpas Pastor and Alvar Ezquerro, 1996; Escartín et al., 2018), and (b) **complex grammatical constructions** that, although not MWEs in the strict sense, contribute semantically or pragmatically to metaphorical meaning.

5.1.1. Multi-Word Expressions (MWEs)

MWEs were identified through a manual, iterative clustering process based on structural patterns described by Corpas Pastor and Alvar Ezquerro (1996), Escartín et al. (2018), and the *Nueva Gramática de la Lengua Española* (NGLE, 2009). The following types were considered relevant for metaphor annotation:

Light Verb Constructions (LVCs). LVCs² were considered for metaphor analysis when the choice of light verb introduced a paradigmatic contrast that shaped the metaphorical framing of the event. Common light verbs include *dar* (to give), *tener* (to have), *tomar* (to take), *recibir* (to receive), and *pasar por* (to go through, to undergo). For instance,

²We take the definition and LVC selection from Alonso Ramos (2004)

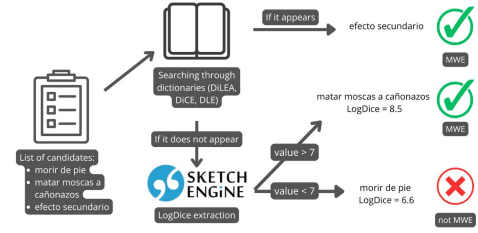


Figure 3: Pipeline MWE curation

in expressions such as *dar miedo* (to scare, lit. to give fear) or *recibir noticias* (to receive news), abstract entities like emotions and information are conceptualized as manipulable objects that can be given or received. Similarly, in *tocar, recibir, or pasar por tratamiento* (to have, to receive, or to go through treatment), the noun *tratamiento* ('treatment') is metaphorically construed as a tangible entity, an impinging external force, or a milestone within a journey. In these cases, LVCs were annotated as part of the metaphorical vehicle because the figurative meaning resides in the interaction between the light verb and its abstract complement, rather than in either element alone.

Fixed and Formulaic Expressions. Highly frequent and context-bound collocations such as greetings, discourse routines, or conventional phrases (e.g., *¿cómo te va?* 'how is it going?' [journey] vs. *¿qué tal estás?* 'how are you?' [neutral]) were annotated when they contributed metaphorical shading to discourse (Dickins, 2020).

Idiomatic Expressions and Collocations. Idioms (e.g., *comerse el coco* 'to overthink', 'to eat one's head', from lit. 'to eat the coconut') were treated as single analytical units due to their semantic opacity. Potential collocations were evaluated for **institutionalization** and **frequency**, and analyzed as single lexical units when they met these criteria. Institutionalization was verified using **DILEA** (Penadés Martínez, 2020), **DICE** (Vincze et al., 2011), and **RAE** resources. Frequency was assessed using the **logDice** metric (Rychlý, 2008), robust to corpus size (Gablasova et al., 2017). Following Pinto et al. (2021), co-occurrences with $\logDice \geq 7$ in the *esTenTen23* corpus (Sketch Engine) (Sketch Engine, 2025) were considered valid phraseological units. See Figure 3 for more details.

Grammatical Constructions Included in the Vehicle. Certain grammatical combinations, while not metaphorical in themselves, modulate the interpretation of metaphors through modality, temporality, or aspectuality. These were included within the metaphorical vehicle but not further analyzed for metaphorical content:

- **Modal verb constructions:** Expressions

with *haber que/de*, *tener que*, *poder*, *deber*, or *querer* (e.g., *debo tomar precauciones* ‘I must take precautions’) adjust the degree of the speaker’s modality, obligation or possibility (Magaña, 2020).

- **Temporal verb constructions:** Forms such as *ir a*, *ir + gerundio*, or *haber + participio* frame metaphorical processes temporally (e.g., *va a pasar por algo parecido* ‘he’s going to go through something similar’).
- **Copulative verbs:** *Ser*, *estar*, and *parecer* frame metaphors by marking states or attributes (*soy una persona fuerte* ‘I am a strong person’, *estoy débil* ‘I am weak’). For instance, *la alimentación y el deporte son aliados* ‘nutrition and exercise are allies’) encodes a direct metaphorical comparison.

5.1.2. Non-MWE Multiword Metaphorical Vehicles

In addition to MWEs, certain multiword combinations were annotated as single metaphorical vehicles when they contributed meaningfully to metaphor construction. Four main cases were identified:

Frame-interaction metaphors. Following Stickles et al. (2016), we included cases where metaphoricality arises from **mismatches between semantic roles** or **cross-frame alignments** rather than individual lexical items, e.g., *lo único que acaba para en mi cabeza es la muerte* ‘the only thing that takes over my head is death’) or *remamos en el mismo barco* ‘we row in the same boat’).

Explicit metaphors. Following the MIPVU framework and extensions to Spanish proposed by Sánchez-Montero et al. (2025a), explicit metaphors and similes containing overt markers or signals (*como si*, *como que*, ‘as if’, ‘as though’) were annotated as multiword vehicles (e.g., *vivo como si mañana no fuera a estar* ‘I live as if tomorrow I would not be here’).

Extended frame metaphors. In some cases, metaphorical content extended coherently across an entire clause or frame, encompassing several lexical items that together instantiate the metaphorical domain (e.g., *es una carrera de obstáculos y hay que ir pasándolos* ‘it is an obstacle race and one must keep overcoming them’).

Hyperbolic expressions. Recurrent hyperboles (*no hay nada que puedas hacer* ‘there’s nothing you can do’, *no pasa nada* ‘it’s fine’, *has hecho todo lo que ha estado en tu mano* ‘you’ve done

everything within your reach’) were annotated as multi-word vehicles when functioning as intensified metaphors rather than literal exaggerations.

5.2. Vocabulary Curation

Given the restricted domain of oncology, many metaphorical expressions were highly recurrent, often used with consistent basic and contextual meanings across posts. To analyze these patterns systematically, we compiled a unified vocabulary. The process began by merging all annotations, which contained an initial list of 2035 candidate items. Subsequent clustering and curation refined this list to a final vocabulary of approximately 800 distinct metaphorical vehicles.

Leveraging the annotated synsets, word senses were clustered according to hypernymy relations to identify shared semantic domains. Once these domains were established, lexical items were reviewed and discussed within their respective clusters. This process facilitated the evaluation of similar semantic behaviors and ensured coherent annotation decisions.

One illustrative case involved weak verbs within the semantic field of movement or transition (e.g., *seguir* ‘to follow/continue’, *comenzar* ‘to begin’, *terminar* ‘to end’, *pasar* ‘to go through’, *pasar por* ‘to undergo’). These verbs are highly grammaticalized, resulting in considerable polysemy and contextual variability, which made it difficult to establish a stable basic sense, leading to the following decisions:

- *Seguir* ‘to follow’ was marked as metaphorical when referring to pursuing or following someone, but not when it meant ‘to continue’.
- *Comenzar* ‘to begin’ and *terminar* ‘to end’ were marked as non-metaphorical.
- *Pasar* and *pasar por* ‘to go through’, ‘to undergo’ were generally marked as metaphorical due to their abstract uses to describe experiences or phases.

Another discussion centered on adjectives describing physical characteristics (e.g., *débil* ‘weak’, *fuerte* ‘strong’, *duro* ‘hard’, *grave* ‘serious’, *agudo* ‘acute’). The team questioned whether the physical-quality sense (e.g., *fuerte* as the ability to exert physical resistance) was more basic than its use to describe a person’s body (e.g., physical strength, muscular build). Since neither could be clearly established as more fundamental without resorting to etymological analysis, which was beyond the scope of this study, these adjectives were only annotated as metaphorical when used to describe mental or emotional processes

(e.g., *una persona fuerte* ‘a strong person’ in the sense of emotional resilience).

Similarly, *controlar* (‘to control’) was marked as metaphorical in expressions such as *controlar el dolor*, *el miedo*, or *la enfermedad* (‘to control pain’, ‘fear’, or ‘the illness’). Although *controlar* has a broad meaning, examining its noun form *control* (‘control’) revealed a more basic sense related to exerting physical power, which supported its metaphorical interpretation in these contexts.

A further case involved *fe* (‘faith’). A distinction was made between *fe* as religious belief (not annotated as metaphorical) and *fe* as confidence or trust in something (annotated as metaphorical), as in *tener fe en el tratamiento* (‘to have faith in the treatment’).

Curation decisions were finalized through a consensus-based review of concordance lines in group sessions, which was essential for resolving borderline cases. To provide transparency, the complete vocabulary file of over 800 metaphorical vehicles, including basic and contextual meanings, is openly available. To the best of our knowledge, this constitutes the largest vocabulary of annotated basic and contextual meanings to date, and the only domain-specific one. Beyond transparency, the explicit annotation of basic meanings also enhances interoperability: since senses are aligned with WordNet (Miller, 1995), the resource can be linked to other semantic frameworks such as Framester (Gangemi, Aldo and Presutti, Valentina and Navigli, Roberto and others, 2017).

Moreover, basic meaning (BM) annotation is a crucial component in several automated metaphor identification methods. However, the lack of manually annotated resources has often led such systems to rely on shortcuts, such as selecting the first dictionary sense, an approach that does not always correspond to the most appropriate basic meaning in language use. By providing a large, systematically annotated BM inventory, this resource can help improve the accuracy and conceptual grounding of future automated metaphor identification efforts. For a general-domain counterpart, see Maudslay and Teufel (2022).

6. Data Analysis

In terms of the annotation itself, we explored two main aspects: (1) the distribution of IAA across different domains to assess which concepts were more subjective or prone to disagreement, and (2) the relationship between metaphorical content and disagreement, based on the hypothesis that words which can be used either metaphorically or literally in the cancer context may lead to greater ambiguity for both manual and automated annotation.

Regarding agreement, as shown in Figure 4, the

most frequent case was that of disagreement. This outcome is not unexpected given the complexity of metaphor annotation and is consistent with previous reports (Sánchez-Montero et al., 2025a). Interestingly, the second most frequent case was that of unanimous agreement, particularly in domains related to emotions. We hypothesize that agreement tends to increase for highly conventionalized metaphors, which are more easily recognized by annotators.

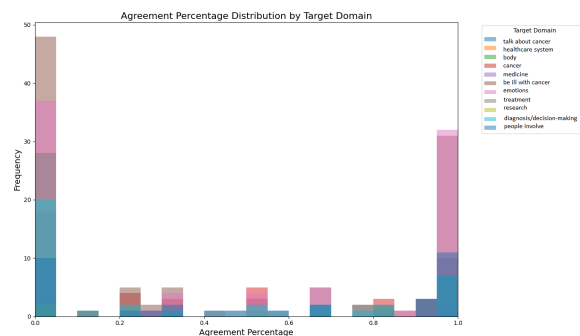


Figure 4: Distribution of agreement across semantic domains.

Secondly, we conducted a thematic analysis to address questions such as: *What are the most metaphorically discussed themes? Are war and journey the only ways of conceptualizing cancer or do other source domains prevail for particular topics?*

To understand how **target domains** (cancer-related topics such as treatment, diagnosis, or emotions) are expressed through **source domains** (the metaphorical images used to represent them, such as war or journey), we leveraged the BM annotations linked to WordNet synsets. These were then queried through the Open Multilingual WordNet (Vossen et al., 2016) to automatically group them into semantic domains.

Figure 5 displays the relationships between cancer-related topics (green nodes) and their metaphorical domains (blue nodes) as a network graph. The network’s central node is the general topic of *being ill with cancer*, which connects to experiences like pain, death, healing, and relapse. This is followed by *emotions*, *people involved* (including emotional support, a key dimension in the forums as users sought shared experiences and advice), *cancer* itself, *diagnosis*, and *decision making*. Less frequently discussed topics, at least metaphorically, included *treatment*, *body*, *research*, and the *healthcare system*. This does not necessarily imply that these topics are less discussed overall, but rather that they appeared less often in this annotated sample of AECC posts.

A closer analysis of the largest nodes, *being ill*

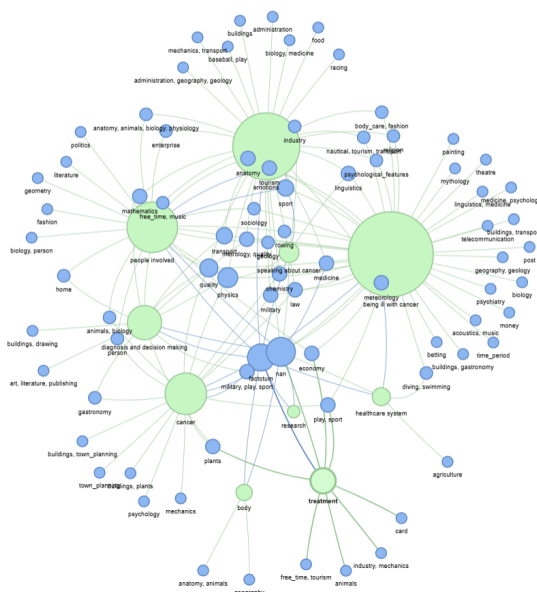


Figure 5: Relations between target (green) and source (blue) metaphorical domains.

with cancer and emotions, reveals that symptoms and side effects were among the most frequently discussed subtopics, followed by life changes, healing, and death. Regarding emotions, beyond general distress, positive emotions such as *optimism* and *hope* were notably prominent, reflecting the supportive tone of the forums, where users actively encourage others seeking solidarity and understanding.



Figure 6: Source domains related to emotions.

In terms of source domains, the **military domain** was indeed central and frequently used across several topics (including emotions, illness, and cancer itself). However, other metaphorical frames were equally or more salient in specific contexts.

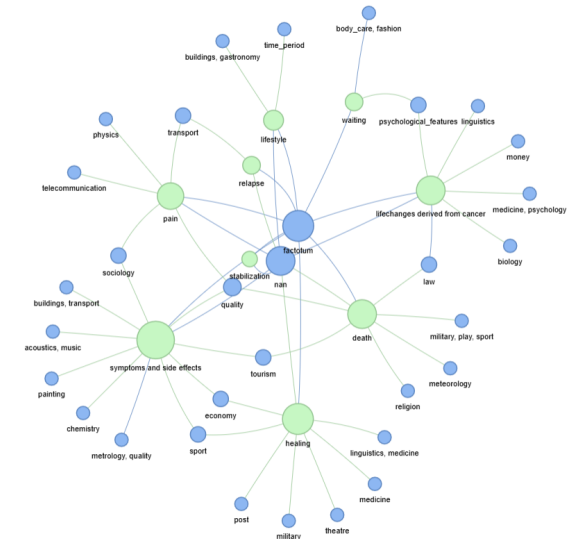


Figure 7: Source domains related to *being ill with cancer*.

For instance, metaphors related to **buildings, construction, and home** appeared recurrently when discussing emotions, emotional support, and even diagnosis. Likewise, **transport and mechanics** metaphors were common in discussions about emotions, cancer, and treatment. These findings demonstrate that Spanish cancer discourse, specifically narratives of cancer experience, draws from a far more diverse metaphorical repertoire than is typically acknowledged in the literature, which opens new avenues for the study of metaphors in Spanish biomedical communication.

7. Conclusions

The paper presents Medical-Flavors-AECC, a dataset of cancer narratives in Spanish extracted from digital forums. We propose a manual annotation framework based on distinguishing basic and contextual meaning, following MIP (Praggle-jaz Group, 2007) and MIPVU (Steen et al., 2010), through the use of various lexical resources, such as WordNet, in different annotation layers.

This resource is intended for the automatic detection of metaphors, especially in the medical field, as well as for the linguistic study of cancer discourse. To this end, we offer an analysis of the metaphorical contents in Spanish cancer narratives, as well as the emotions involved in the production of these metaphors and the degree of lexicalization across different conceptual domains.

This research has significant social and humanistic impact. The study's results can help physicians and healthcare personnel improve their commu-

nication strategies when discussing the disease. At the same time, our results provide psychologists with an empirical basis for investigating how metaphors shape the perception of illness.

As a future work, we will seek to test alternative forms of annotation frameworks that allow for more extensive research into the subjective scope of metaphor and lexicalization processes in cancer narratives. Another area of current development is the annotation of a larger dataset using LLMs, including hybrid labeling approaches that combine human-AI collaboration. Future work will involve extending the annotation scheme with multiclass labels to distinguish between different metaphor types. This will enable a more fine-grained analysis of the specific metaphors characteristic of each conceptual domain.

8. Data Statement

We follow the guidelines specified by (Bender and Friedman, 2018) for creating a Data Statement to mitigate bias and ensure transparency in collection and annotation of the data.

A. Curation Rationale Posts were collected from the public forums of the Spanish Association Against Cancer (AECC). Given the association’s prominence and its broad coverage of patients, relatives, and experiences across time, the AECC forums offered an ethically appropriate and representative source of cancer discourse in Spanish.

B. Language Variety The publications are written in Peninsular (European) Spanish. Orthographic and lexical variation was preserved during the annotation process to maintain linguistic authenticity, as the orthography employed in this type of public forum often deviates from normative conventions. Non-textual content such as images or links was removed, while the original punctuation, capitalization, and emoticons were retained.

C. Forums Authors Demographic The authors of the posts are patients, relatives, and caregivers participating in AECC online forums between 2010 and 2024. Since all forums contributions were publicly posted and pseudonymized, no personally identifiable or demographic information (e.g., age, gender, location) was collected. We expect diversity in socioeconomic background, education level, and regional origin among contributors since the AECC’s broad reach in Spain.

D. Annotator Demographic The annotation team comprised five independent annotators, all native Spanish speakers previously trained in the

Categories	Data
Sex	2 female
	3 male
Native Language	Spanish
Nationality	4 Spanish
	1 Mexican
Residence	Spain and Mexico
Education level	University

Table 3: Annotator demographic

Metaphor Identification Procedure (MIP) and with experience in metaphor annotation. Four were trained linguists, and one had a background in Natural Language Processing (NLP), Table 3 summarizes their profiles. All annotators provided informed consent, and the annotation process adhered to the principles of the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979).

E. Speech Situation The forums posts are about personal experiences with cancer, including diagnosis, treatment, recovery, and emotional support. Topics vary from medical advice and coping strategies to interpersonal communication and testimonies. Posts were written asynchronously in a semi-formal or non-formal tone, frequently used in online patient support environments.

F. Text Characteristics The dataset includes 7,964 forum posts (5,641,005 words³), of which 74 posts (88,196 words) were manually annotated for metaphor use. Posts cover three main topics: breast cancer, prostate cancer, and personal testimonies. Texts were pseudonymized before annotation, with user mentions, personal names, and identifiable references removed or replaced following GDPR-compliant pseudonymization protocols (Parlamento Europeo y Consejo de la Unión Europea, 2016; Búrdalo, 2025; López-García et al., 2023). Additionally, each of the 74 posts includes comments that respond directly to the main topic introduced in the main post.

G. Recording Quality The text was processed into UTF-8 plain text format, ensuring consistency in encoding and tokenization. We scraped the posts from the AECC online forums. Descriptive statistics are presented in Table 4.

H. Ethical Considerations The data collection and annotation procedures were reviewed and approved by three independent entities: the AECC,

³Text splitting into words was done automatically by INCEpTION software.

Metric	Total	Annotated
Number of words	5,641,005	83,316
Number of posts	7,964	72
Number of users	2,976	442
Avg. words per post	708.30	
Avg. interactions (main author)	2.58	
Avg. interactions (comments)	12.20	
	Breast cancer = 504.5	
	Prostate cancer = 784.7	
	Testimonies = 749.8	
Avg. words per topic		
Covered years	2010-2024	

Table 4: Descriptive statistics of Medical-FLAVORS AECC full corpus and the presented annotated dataset.

the University of Zaragoza, and the EU project funding the research. All posts used were publicly available and pseudonymized to protect participants' identities. Annotators were fully briefed about the study's purpose and informed that the content could be emotionally sensitive. They were free to withdraw from participation at any point. Finally, this procedure followed the General Data Protection Regulation of the European Union (GDPR; Regulation (EU) 2016/679) ([Parlamento Europeo y Consejo de la Unión Europea, 2016](#)), established guidelines for anonymizing Spanish medical texts ([Búrdalo, 2025](#)), and existing anonymization tools for Spanish ([López-García et al., 2023](#)).

9. Limitations

We acknowledge some limitations in our study. First, although the dataset comprises over 83,000 manually annotated words, inter-annotator agreement on the curation could not be quantified, and the annotations result from consensus between two annotators and are then curated by a single curator.

Second, the calculation of metaphorical density relied in part on automatic lemmatization using SpaCy. During manual inspection, several lemmatization errors were identified, particularly involving irregular verb forms and clitic pronouns, which may have affected the precision of metaphor frequency estimates. Consequently, the metaphoricity percentages reported should be interpreted as approximate rather than exact.

Third, not all lexical items in the vocabulary could be assigned a WordNet synset. In some cases, no synset adequately captured the sense of a word as used in the dataset. These items were therefore excluded from the semantic domain analysis. Moreover, while WordNet domains are suitable for nominal concepts, many verbal synsets are assigned to the generic domain *factotum*, limiting their

discriminative value. Future work should explore frame-based mappings (e.g., FrameNet) or manual domain annotation to improve coverage and granularity for verbs and other non-nominal items.

Finally, subsequent iterations of this resource will benefit from integrating domain-adapted lemmatization models and incorporating peer-reviewed quality control to further enhance reliability and reproducibility.

10. Acknowledgements

This paper has been supported by PA-PIIT project IG-400325, by the I+D+i projects PID2024-159530OB-I00 (funded by MCIN/AEI/10.13039/501100011033), the EU research and innovation program HORIZON Europe in the “4D PICTURE” project under grant agreement 101057332 and UZ2024-IyA-02 (funded by Univ. Zaragoza), and by DGA Government predoctoral fellowship.

11. Bibliographical References

- Margarita Alonso Ramos. 2004. *Las construcciones con verbo de apoyo*. Visor Libros.
- Asociación de Academias de la Lengua Española. 2023. *Diccionarios de la lengua española*. Accessed: 2025-10-19.
- Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Mayla Boguslav and Kevin Bretonnel Cohen. 2017. Inter-annotator agreement and the upper limit on machine performance: evidence from biomedical natural language processing. In *MEDINFO 2017: Precision Healthcare through Informatics*, pages 298–302. IOS Press.
- Luis Búrdalo. 2025. *D4.2 final harpocrates demonstrators - appendix i*. Consultado: 2025-10-19.
- Lynne Cameron. 2007. The affective discourse dynamics of metaphor clustering. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies*, (53):041–062.
- David Casarett, Amy Pickard, Jessica M. Fishman, Stewart C. Alexander, Robert M. Arnold, Kathryn I. Pollak, and James A. Tulsky. 2010. *Can metaphors and analogies improve communication with seriously ill patients?* *Journal of Palliative Medicine*, 13(3):255–260. PMID: 19922170.

- Gloria Corpas Pastor and Manuel Alvar Ezquerra. 1996. *Manual de fraseología española*. Gredos.
- Laurence De Backer, Renata Enghels, and Patrick Goethals. 2023. Metaphor analysis meets lexical strings: finetuning the metaphor identification procedure for quantitative semantic analyses. *Frontiers in Psychology*, 14:1214699.
- James Dickins. 2020. An ontology for collocations, formulaic sequences, multiword expressions, compounds, phrasal verbs, idioms and proverbs. *Linguistica Online*, 23:29–72.
- Ellen Dodge, Jisup Hong, and Elise Stickles. 2015. [MetaNet: Deep semantic automatic metaphor analysis](#). In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49, Denver, Colorado. Association for Computational Linguistics.
- Carla Parra Escartín, Almudena Nevado Llopis, and Eoghan Sánchez Martínez. 2018. Spanish multiword expressions: Looking for a taxonomy. *Multiword expressions*, page 271.
- Nada Fadul, Ahmed F. Elsayem, J. Lynn Palmer, Egidio G Del Fabbro, Kay Swint, Zhijun Li, Valerie A Poulter, and Eduardo Bruera. 2009. Supportive versus palliative care: What's in a name? *Cancer*, 115.
- Dana Gablasova, Vaclav Brezina, and Tony McEnery. 2017. [Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence](#). *Language Learning*, 67(S1):155–179.
- Gangemi, Aldo and Presutti, Valentina and Navigli, Roberto and others. 2017. [Framester: A wide-coverage linguistically and cognitively grounded knowledge graph](#). Accessed: 2025-10-20.
- Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artif. Intell. Rev.*, 56(S2):1829–1895.
- Anna W. Gustafsson, Charlotte Hommerberg, and Anna Sandgren. 2019. Coping by metaphors: the versatile function of metaphors in blogs about living with advanced cancer. *Medical Humanities*, 46:267 – 277.
- George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. The University of Chicago Press.
- Mark Jordan Landau, Linda D. Cameron, Jamie Arndt, W. Hamilton, Trevor James Swanson, and Michael N. Bultmann. 2019. Beneath the surface: Abstract construal mindset increases receptivity to metaphors in health communications. *Social cognition*, 37 3:314–340.
- Yufeng Liu, Elena Semino, Judith Rietjens, and Sheila Payne. 2024. [Cancer experience in metaphors: patients, carers, professionals, students – a scoping review](#). *BMJ Supportive & Palliative Care*, 14(e3):e2366–e2376.
- Guillermo López-García, Francisco J. Moreno-Barea, Héctor Mesa, José M. Jerez, Nuria Ribelles, Emilio Alba, and Francisco J. Veredas. 2023. [Named entity recognition for de-identifying real-world health records in spanish](#). In *Proceedings of the 23rd International Conference on Computational Science (ICCS 2023)*, volume 14073 of *Lecture Notes in Computer Science*, pages 172–183. Springer. Consultado: 2025-10-19.
- Dalia Magaña. 2020. Local voices on health care communication issues and insights on latino cultural constructs. *Hispanic Journal of Behavioral Sciences*, 42(3):300–323.
- Rowan Hall Maudslay and Simone Teufel. 2022. Metaphorical polysemy detection: Conventional metaphor meets word sense disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 65–77.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. The belmont report: Ethical principles and guidelines for the protection of human subjects of research. Technical report, U.S. Department of Health, Education, and Welfare, Washington, D.C. Publication No. OS 78-0012.
- Parlamento Europeo y Consejo de la Unión Europea. 2016. [Reglamento \(ue\) 2016/679 del parlamento europeo y del consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos](#). Consultado: 2025-10-19.
- Inmaculada Penadés Martínez. 2020. [Diccionario de locuciones idiomáticas del español actual](#). Accessed: 2025-10-19.
- Ana P. M. Pinheiro, Rachel H. Pocock, Margie D. Dixon, Walid L Shaib, Suresh S. Ramalingam, and Rebecca D. Pentz. 2017. Using metaphors to explain molecular testing to cancer patients. *The oncologist*, 22 4:445–449.
- Paula Tavares Pinto, Geraint Paul Rees, and Ana Frankenberg-Garcia. 2021. [Identifying collocation issues in english l2 research article writing](#). In Maggie Charles and Ana Frankenberg-Garcia,

- editors, *Corpora in ESP/EAP Writing Instruction: Preparation, Exploitation, Analysis*, chapter 7, page 24. Routledge.
- Lucia Pitarch, Emma Angles-Herrero, Yufeng Liu, Daisy Monika Lal, Jorge Gracia, Paul Rayson, and Judith Rietjens. 2024. [Medical-FLAVORS: A figurative language and vocabulary open repository for Spanish in the medical domain](#). In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 109–114, Torino, Italia. ELRA and ICCL.
- Pragglejaz Group. 2007. [MIP: A method for identifying metaphorically used words in discourse](#). *Metaphor and Symbol*, 22(1):1–39.
- Real Academia Española. 2023. [Diccionario de la lengua española, 23.ª edición](#). Accessed: 2025-10-19.
- Real Academia Española and Asociación de Academias de la Lengua Española. 2009. *Nueva Gramática de la Lengua Española*. Espasa Libros / RAE-ASALE, Madrid, Spain. Accessed: 2025-10-19.
- Pavel Rychlý. 2008. A lexicographer-friendly association score. In *RASLAN*, pages 6–9.
- Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra. 2025a. [Disagreement in metaphor annotation of Mexican Spanish science tweets](#). In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 155–164, Abu Dhabi, UAE. International Committee on Computational Linguistics.
- Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra. 2025b. [Prompting metaphoricity: Soft labeling with large language models in popular communication of science tweets in Spanish](#). In *Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, pages 45–56, Vienna, Austria. Association for Computational Linguistics.
- Elena Semino and Zsófia Demjen. 2017. The cancer card: Metaphor, intimacy, and humor in online interactions about the experience of cancer. *Metaphor: Embodied cognition and discourse*, page 181.
- Elena Semino, Zsófia Demjén, Andrew Hardie, Sheila Payne, and Paul Rayson. 2017. *Metaphor, cancer and the end of life: A corpus-based study*. Routledge.
- Sketch Engine. 2025. [Estenten corpus](#). Accessed: 2025-10-20.
- Sociedad Española de Oncología Médica. 2025a. [Cáncer de mama](#). Accessed: 2025-10-19.
- Sociedad Española de Oncología Médica. 2025b. [Los avances en cáncer de próstata](#). Accessed: 2025-10-19.
- Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company.
- Elise Stickles, Oana David, and Eve Sweetser. 2016. Grammatical constructions, frame structure, and metonymy: Their contributions to metaphor computation. In *Proceedings of the 11th meeting of the High Desert Linguistics Society*, pages 317–345. High Desert Linguistics Society Albuquerque, NM.
- Orsolya Vincze, Elena Mosqueira, and Manuel Alonso Ramos. 2011. An online collocation dictionary of spanish. In *Proceedings of the 5th International Conference on Meaning-Text Theory*, pages 275–286, Barcelona, Spain.
- Piek Vossen, Francis Bond, and John McCrae. 2016. [Toward a truly multilingual GlobalWordnet grid](#). In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 424–431, Bucharest, Romania. Global Wordnet Association.
- World Health Organization. 2025. [Breast cancer](#). Accessed: 2025-10-19.

A Synthetic Conversational Dataset for Type 2 Diabetes Management

Stergios Ntanavaras¹, Maaïke H. T. de Boer², Piek Vossen¹

¹ Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands

² TNO, 3769 DE Soesterberg, The Netherlands

s.ntanavaras@student.vu.nl, maaïke.deboer@tno.nl, piek.vossen@vu.nl

Abstract

Access to real patient-doctor conversations in the medical domain is often restricted due to privacy concerns, making it difficult to build robust conversational AI systems. To address this, we present a novel methodology for generating a high-quality synthetic dataset designed for conversational triple extraction in Type 2 Diabetes management. Using structured prompting with GPT-4, we generated 16 demographically and medically diverse diabetic personas, and 256 multi-turn conversations between these personas and a caretaker agent, simulating realistic and context-rich interactions. The conversations incorporate critical properties such as personalization, empathy, contextual awareness, and medically grounded advice, as validated through both LLM-based and human expert evaluations. These synthetic conversations are further annotated with Subject-Predicate-Object (SPO) labels at the token level, integrating both manual and LLM-automated methods, forming the foundation for downstream tasks like triple extraction. Our work demonstrates the feasibility of using generative AI to simulate healthcare conversations at scale, offering a solution for data-scarce domains.

Keywords: Synthetic Data Generation, Conversational AI, Dialogue Annotation, Conversational Triple Extraction, Type 2 Diabetes Management

1. Introduction

The increasing burden of chronic diseases such as Type 2 Diabetes poses a significant challenge to healthcare systems worldwide, particularly in aging populations (Changizi and Kaveh, 2017). Traditional healthcare infrastructures often struggle to provide personalized and accessible care on a scale (Wagner et al., 1996), creating an urgent need for innovative solutions that can support both patients and healthcare professionals in managing these complex conditions.

Hybrid Intelligence (HI) and conversational agents offer a promising way to support chronic disease management, including Type 2 Diabetes, by making care more personalized and accessible (Laranjo et al., 2018; Dudzik et al., 2024). They can provide around-the-clock availability, consistent information delivery, and scalable patient engagement, while reducing the burden on healthcare professionals. However, developing and evaluating such agents requires access to high-quality domain-specific datasets that reflect the nuances of real patient-agent interactions.

Unfortunately, several critical gaps exist in the current landscape of healthcare dialogue datasets. First, especially for chronic diseases like Type 2 Diabetes, data privacy regulations restrict access to authentic patient dialogues, resulting in a scarcity of annotated conversational data. Second, most existing datasets are general-purpose and do not address the specific informational and behavioral

needs associated with managing Type 2 Diabetes (Peerbolte et al., 2025). Lastly, most of the available dialogue resources are based on human-human interactions, offering limited support for training and evaluating agent-driven communication. Together, these limitations present a critical bottleneck in the advancement of language technologies for medical settings.

To overcome this challenge, this paper presents a methodology for generating a high-quality synthetic conversational dataset tailored to the context of Type 2 Diabetes management. These synthetic conversations simulate realistic interactions between patients and a caretaker agent, providing the necessary foundation for downstream tasks such as triple extraction and knowledge graph enrichment. More specifically, the dataset is designed to train and benchmark models that extract Subject-Predicate-Object (SPO) triples from patient utterances, which are then used to populate and maintain a User Knowledge Graph within a broader HI system for personalized lifestyle support.

We propose a methodology for constructing such a dataset using structured prompting with detailed task instructions and large language models (LLMs). Our approach combines persona-driven scenario creation with dialogue generation, guided by design principles from healthcare communication and conversational AI. We describe how we created culturally grounded patient personas, generated multi-turn dialogues, and ensured diversity and realism across interactions.

This work is situated within the CHIP initiative (den Hengst et al., 2026), a collaborative effort by Dutch institutions to develop an HI system for Type 2 Diabetes management. CHIP is a modular, microservice-based research platform that integrates dialogue processing, information extraction, knowledge graphs, and reasoning to support lifestyle changes in patients. Within CHIP’s architecture, conversational interactions are first processed by a knowledge extractor, which converts free-form dialogue text into SPO triples. These triples are stored in a User Knowledge Graph alongside a Domain Knowledge Graph of medical knowledge. A reasoning component then consults both graphs to tailor interventions to individual patient contexts and goals. The synthetic dialogues and conversational triple extraction methods presented in this paper directly support the knowledge extraction step in this pipeline, providing annotated training and evaluation data for models that must operate on realistic, domain-specific patient–agent exchanges. By generating this resource synthetically, we circumvent the privacy constraints that would otherwise prevent the use of real patient dialogues for model development.

To support transparency and reproducibility, all diabetes personas, synthetic conversations, and their corresponding annotations are available in the accompanying repository ¹.

2. Related Work

Synthetic dataset generation is increasingly seen as a practical way to address data scarcity in conversational AI, especially in domains with privacy or access constraints (Chavda and Bhattacharyya; Ramesh et al., 2024). Early studies explored data augmentation techniques, using rule-based transformations or back-translation to expand existing corpora, but these approaches produced limited linguistic and contextual diversity (Soudani et al., 2023).

With the advent of large language models (LLMs), researchers have shifted toward generating full dialogues synthetically. For instance, Honovich et al. (Honovich et al., 2022) demonstrated how prompt-based models can be leveraged to automatically generate diverse synthetic training conversations that capture richer context, intent, and diversity while reducing dependency on costly or restricted real-world data.

In clinical and healthcare dialogue modeling, a smaller but growing body of work has begun producing synthetic resources. SynDial (Das et al.,

2024) generates synthetic patient–physician dialogues from clinical notes and uses iterative refinement to maintain factuality. NoteChat (Wang et al., 2024) produces doctor-patient conversations conditioned on clinical notes for better documentation and note generation performance. MedSynth (Mianroodi et al., 2025) provides over 10,000 dialogue-note pairs across many disease codes, facilitating both Dialogue-to-Note and Note-to-DIALOGUE tasks. Bedrick et al. (Bedrick et al., 2025) offer a typology of existing synthetic datasets for clinical dialogues, highlighting how synthetic methods vary in realism, domain specificity, and use case.

Beyond task-oriented generation, recent research has emphasized the role of social and personal dimensions in synthetic dialogues. Santamaría et al. (Santamaría et al., 2023) argue that open-domain dialogue datasets should not only encode factual content but also capture personal viewpoints and shared memories over time. Their work highlights the need for structured representations that track factoids and perspectives across long-term interactions, which can inspire the development of more transparent and socially attuned conversational agents. This perspective resonates with the challenges in healthcare, where patient narratives are inherently personal and context-dependent.

Despite these advances, synthetic datasets tailored for downstream tasks such as conversational triple extraction remain scarce. This is especially true for resources that include Subject-Predicate-Object (SPO) annotations, multi-turn patient–agent dialogues, and domain grounding for Type 2 Diabetes. Our dataset and study address this gap by combining synthetic dialogue generation, persona grounding, comprehensive SPO annotation, and evaluation in a knowledge-graph-oriented pipeline for CHIP.

3. Methodology

To ensure practical utility and align with the needs identified in Section 2, we target three constraints: limited access to real dialogues due to privacy, lack of domain-specific resources for Type 2 Diabetes, and the need to model agent-driven interactions. We define design requirements to guide the creation of the synthetic conversational dataset and to align with the CHIP initiative.

The dataset was designed to meet the following requirements:

- All conversations focus on Type 2 Diabetes management, covering topics such as symptoms, treatment options, lifestyle changes, and patient concerns.
- Dialogues should reflect *patient–agent* setting rather than *human–human* interaction.

¹<https://github.com/stergios97/synthetic-conversational-data-diabetes-management>

- Each dialogue contains 4–6 exchanges with natural turn-taking; openings are roughly balanced between patient-start and agent-start to avoid initiation bias.
- Diversity across intents, scenarios, and personas, with coverage of all defined profiles.
- Only fully synthetic content, with no real patient data involved, in order to comply with privacy standards and ethical research practices.

The data generation process consists of three steps: creating diverse patient personas, generating multi-turn dialogues with a caretaker agent, and annotating conversations with Subject–Predicate–Object (SPO) roles. The result is a dataset of fully annotated conversational sentences, with each step described in the following sections.

3.1. Persona Generation

For making the dataset more natural, we started by creating diverse personas, each representing a unique profile of a Type 2 Diabetes patient. These personas were carefully designed to capture a wide range of demographic traits, medical histories, and treatment adherence behaviors, providing a detailed representation of the patient population. The selection of specific demographic attributes for the creation of personas was informed by robust epidemiological and sociocultural research, along with insights from domain experts, ensuring that the personas accurately reflect the diversity and nuances of the population of patients with Type 2 diabetes in the Netherlands.

To construct realistic patient personas, we grounded our trait selection in epidemiological and demographic data from Dutch public health sources. Age values were chosen around the national mean of 65.9 years for Type 2 Diabetes outpatients, as reported by the Dutch Pediatric and Adult Registry of Diabetes (DPARD) (Bak et al., 2021). We selected five representative age groups (55, 60, 65, 70, 75) to reflect the elderly population most impacted by the disease. The ethnicity was chosen based on prevalence statistics from StatLine, the Dutch Central Bureau of Statistics (CBS) (Centraal Bureau voor de Statistiek, 2024) and supporting studies (Ujcic-Voortman et al., 2009), which show elevated Type 2 Diabetes rates among Moroccan (7.09%), Surinamese (6.15%), and Turkish (5.95%) populations, compared to Dutch individuals (1.99%). Therefore, these ethnic groups were selected for inclusion, given both their medical relevance and their demographic presence in the Netherlands. To ensure cultural authenticity, we assigned each persona a name commonly associated with their ethnic background, as verified through multiple public sources². Moroccan personas were

named Mohammed, Abdullah, Aicha, and Fatima; Surinamese: Rudolf, Johan, Julia, and Ingrid; Turkish: Ali, Mehmet, Ayşe, and Fatma; Dutch: Pieter, Jan, Johanna, and Maria.

Using these attributes, all possible combinations were generated, respecting gender-specific naming conventions to ensure realism and cultural appropriateness. From these, 16 unique profiles were carefully selected to achieve a balanced and manageable set that covers the full range of age, gender, education, and cultural background while ensuring equal representation across gender and ethnicity. The number 16 was chosen as it provides sufficient diversity without overwhelming the annotation process. All selected profiles are presented in Table 1.

Name	Gender	Ethnicity	Age
Jan	Male	Dutch	70
Maria	Female	Dutch	60
Abdullah	Male	Moroccan	70
Mohammed	Male	Moroccan	75
Aicha	Female	Moroccan	75
Julia	Female	Surinamese	55
Ingrid	Female	Surinamese	75
Johan	Male	Surinamese	60
Fatima	Female	Moroccan	65
Rudolf	Male	Surinamese	55
Ayşe	Female	Turkish	70
Pieter	Male	Dutch	75
Johanna	Female	Dutch	75
Fatma	Female	Turkish	65
Ali	Male	Turkish	60
Mehmet	Male	Turkish	65

Table 1: Selected Profiles

3.2. Prompts for Persona Generation

Once the core demographic attributes were selected, we generated detailed biographies for each persona using structured prompting with detailed task instructions. To guide the model in producing realistic and context-rich outputs, we designed manual prompts that elicited narrative descriptions that include key dimensions of diabetes care. These included personal and demographic backgrounds, physical features, psychological traits, professional and educational experiences, social interactions, daily routines, healthcare interactions, economic factors, communication styles, and technological engagements.

The generation was carried out using GPT-4³, identified as the most proficient model within the GPT series according to a report from OPENAI (OpenAI, 2024). To generate detailed and informative persona descriptions, the `temperature`

within each ethnic group, as verified across multiple public online sources.

³OpenAI gpt-4 (June 2023 release, gpt-4-0613)

²Names were chosen based on their common usage

parameter was set to 0.9 to optimize for creativity, whilst the `max_tokens` was left at its default value of 4096 from the API configuration, ensuring that the descriptions were comprehensive and rich in detail.

The resulting descriptions captured interrelated aspects of each persona’s life that could influence their interaction with a conversational healthcare agent. For instance, some personas were digitally literate and health-conscious, while others had limited access to care or expressed mistrust in medical institutions.

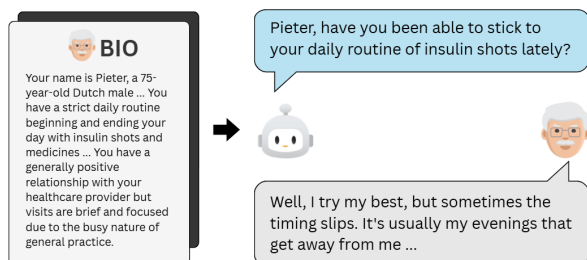


Figure 1: Example of conversation generation from a persona profile. The biography of “Pieter” (75-year-old Dutch male with Type 2 Diabetes) is used to generate a sample dialogue with the caretaker agent.

3.3. Conversation Generation

Building on the finalized personas, we generated synthetic patient-agent dialogues to simulate realistic Type 2 Diabetes management scenarios. This transition from persona definition to conversation generation is illustrated in Figure 1. For each of the 16 personas, we produced 16 conversations. This number provided adequate topical diversity per persona while keeping manual checks and subsequent annotation/evaluation feasible and balanced across the corpus (equal gender/ethnicity coverage and manageable size). Conversations were evenly divided between patient- and agent-initiated openings, and each dialogue was restricted to 4 to 6 exchanges with natural turn-taking and internal coherence.

The final dataset consists of 256 dialogues covering a wide range of diabetes-related topics, including medication adherence, lifestyle adjustments, emotional support, and access to care. These conversations provide a robust foundation for training and evaluating conversational systems focused on personalized health communication. Details of the prompting setup are given in Section 3.4. Below are two examples of the conversations generated.

Patient-initiated example (“Abdullah”, Table 1):

- P:** I’m worried about fasting for Ramadan with my diabetes. Is it possible without harming my health?
- A:** Yes, Abdullah, fasting during Ramadan is possible with diabetes, but it requires careful planning. You should consult your healthcare provider to adjust your medication and meal times according to your fasting schedule.

[conversation continues...]

Agent-initiated example (“Maria”, Table 1):

- A:** Maria, how has your blood sugar control been over the past week? Do you notice any fluctuations?
- P:** Well, lately I’ve noticed that my blood sugar tends to increase significantly after my dinner meal.
- A:** I see. Can you share what a typical dinner looks like for you?
- P:** Usually, it’s a lean protein like chicken, and I try to fill half my plate with vegetables.

[conversation continues...]

3.4. Prompts for Conversation Generation

The generation of realistic and varied dialogues required carefully designed prompts tailored to the roles of the patient and the agent, as well as a coordinating system prompt. Each component was constructed to ensure coherent, context-sensitive exchanges grounded in the persona characteristics.

The *system prompt* defined the scope and structure of each dialogue, instructing the model to produce exactly 4 to 6 exchanges per conversation. To enforce this constraint and improve adherence, we applied strategies such as capitalization and repetition of critical instructions. These techniques enhanced the model’s compliance with formatting and interaction rules during generation.

The *agent prompt* was crafted to encourage empathetic, adaptive, and informative behavior. Drawing on healthcare communication research, we included directives for using friendly and supportive language, engaging in small talk, and adapting responses to the patient’s emotional and informational needs. The agent was also instructed to avoid medical jargon, maintain conversational tone, and gather information incrementally, asking one question at a time to avoid overwhelming the patient.

For the *patient prompt*, we injected each persona biography directly into the input to simulate realistic behavior aligned with their background, beliefs, and

Overview		Dialogue Acts	
Number of dialogues	256	Statements	77.8%
Total utterances	2,027	Comments	7.5%
Total tokens	52,019	Opinions	7.3%
Unique word types	5,337	Positive answers	2.8%
Avg. utt./dialogue	7.9	Commands	2.8%
		Back-channeling	0.7%
Emotions			
Neutral	785	Curiosity	432
Gratitude	211	Approval	142
Fear	33	Disappointment	11
Sadness	19		

Table 2: Dataset statistics, dialogue acts, and emotions.

communication style. This allowed each dialogue to reflect the unique characteristics of the persona, including their health concerns, digital literacy, or trust in medical advice.

We experimented with both zero-shot and few-shot prompting approaches. Zero-shot prompts often failed to produce the required dialogue length or diversity. In contrast, few-shot prompting significantly improved both structure and variety. As a result, we adopted a few-shot format using four curated examples per prompt to guide the model toward generating coherent, contextually rich conversations.

All conversations were generated via the OpenAI GPT-4 model (same version as noted earlier); the generation parameters were set to a `temperature` of 0.9 to encourage linguistic variation and a `max_tokens` of 300 to keep outputs concise.

All prompts used in the generation process are made available in the accompanying repository ¹.

4. Dataset Analysis and Evaluation

4.1. Statistical Analysis

Table 2 summarizes key dataset statistics, dialogue act distributions, and emotion annotations. The synthetic dataset comprises 256 multi-turn patient-agent conversations related to Type 2 Diabetes management, totaling 2,027 utterances and 52,019 tokens, with a vocabulary of 5,337 unique word types. On average, each dialogue contains 7.9 utterances, demonstrating a moderately complex interaction structure.

Dialogue acts were annotated using a method consistent with the approach described by Santamaría et al. (Santamaría et al., 2023), implemented with the MIDAS dialogue act classifier (Yu and Yu, 2019). Specifically, each utterance was passed through a Transformers text classification pipeline loading a MIDAS-based XLM-RoBERTa

model, and the top-ranked dialogue act label was assigned. Analysis of dialogue acts reveals that statements dominate (77.8%), followed by comments (7.5%), opinions (7.3%), and positive answers (2.8%). Commands and back-channeling are less frequent, at 2.8% and 0.7% respectively. This distribution reflects a dataset focused on information exchange and patient education, with limited small-talk or filler content.

Emotion annotations were derived using the same pipeline-based approach, this time loading the GoEmotions BERT model (Demszky et al., 2020) and selecting the top-ranked emotion label per utterance. The results indicate that neutral utterances are the most prevalent (785 instances), followed by curiosity (432), gratitude (211), and approval (142). Negative emotions such as fear (33), disappointment (11), and sadness (19) appear, but are relatively rare. These results show that while patient concerns are captured, the conversational tone remains largely informative and supportive.

The dataset demonstrates high lexical diversity with a type-token ratio (TTR) of 0.10 (5,337 unique types over 52,019 tokens). Subjects are mostly first- and second-person pronouns (“I”, “you”), aligning with the interactive nature of patient-agent dialogues. Frequent predicates include “be”, “try”, “feel”, and “make”, indicating a strong focus on advice, recommendations, and patient states.

Utterance lengths vary substantially, with character counts ranging from 37 to 352. However, most responses fall within 80–150 characters, balancing naturalness and informativeness. Additionally, there is minimal reliance on very short utterances (e.g., “Yes”, “No”), which supports richer patient-agent interactions.

Compared to the open-domain dialogue datasets analyzed by Santamaría et al. (Santamaría et al., 2023) and discussed in Section 2, our dataset differs substantially in domain specificity, conversational structure, and interactional goals. While open-domain resources often emphasize personal perspectives, emotional variability, and human-human social bonding, our dataset is designed to simulate patient-agent interactions focused on fact-based, medically accurate exchanges. Conversations are shorter than narrative-rich corpora like PEDC (Eisenberg and Sheriff, 2020) but exhibit greater linguistic variation and topical specificity than most task-oriented datasets. This unique combination of healthcare-focused, multi-turn dialogues with annotated emotional and linguistic features makes the dataset well-suited for training and evaluating conversational agents in Type 2 Diabetes management, addressing limitations in existing resources.

4.2. Persona Evaluation

To ensure the generated personas were both realistic and relevant for the diabetes care domain, all profiles underwent a manual evaluation and refinement process in collaboration with domain experts—researchers working on diabetes lifestyle support with a background in health communication and conversational AI. This review focused on validating the coherence, plausibility, and representativeness of each profile in relation to the challenges faced by patients managing Type 2 Diabetes.

During the review, particular attention was given to aligning personas with the broader healthcare challenges motivating CHIP. For instance, initial versions of the generated biographies included several well-off individuals with strong healthcare support networks. To better reflect the systemic gaps that CHIP aims to address, these profiles were revised to incorporate traits such as financial hardship, limited access to care, and lower health or digital literacy.

This expert-guided refinement ensured that each persona embodied characteristics that would meaningfully test the capacity of conversational agents to respond to real-world complexities in patient communication and care personalization. The final set of personas thus captures a wide spectrum of needs, barriers, and behaviors, making them a robust foundation for generating diverse and clinically relevant synthetic dialogues.

4.3. Conversation Evaluation

To assess the linguistic quality of the conversations generated, we performed an LLM-based and a human-based evaluation using a custom questionnaire. Both approaches focused on four core dimensions of language quality: fluency, naturalness, clarity, and structure. These dimensions were selected to reflect how effectively the dialogues simulate natural and coherent patient–agent communication.

4.3.1. LLM-based Evaluation

All 256 generated dialogues were included in the LLM-based evaluation process following the G-Eval framework (Liu et al., 2023). Each dialogue was holistically assessed on each of the four criteria mentioned above using GPT-4o⁴. The model was instructed to assign a probability distribution on a 5-point Likert scale (Likert, 1932) based on a detailed rubric per dimension. Then, we computed two key metrics from the outputs: 1. the expected score: a weighted average representing the most likely rating, and 2. the entropy: a measure of

confidence (lower entropy indicating higher confidence). To ensure reproducibility, the evaluation was conducted with `temperature` set to 0.0, making the process deterministic. Table 3 summarizes the average scores and their variability (standard deviations) across dialogues.

Dimension	Avg. Score ($\pm$ SD)	Avg. Entropy ($\pm$ SD)
Fluency	4.70 $\pm$ 0.14	0.64 $\pm$ 0.19
Naturalness	4.45 $\pm$ 0.09	0.91 $\pm$ 0.04
Clarity	4.47 $\pm$ 0.05	0.91 $\pm$ 0.02
Structure	4.14 $\pm$ 0.12	0.86 $\pm$ 0.10

Table 3: GPT-4o evaluation results showing mean $\pm$ SD of scores (1–5 scale) and entropy across 256 dialogues.

As shown in Table 3, the synthetic dialogues achieved consistently high scores across all four dimensions, with particularly strong performance in fluency (4.70 $\pm$ 0.14) and clarity (4.47 $\pm$ 0.05). These results suggest that the model consistently judged the conversations as grammatically correct, clear, and easy to follow. The small standard deviations indicate that these high scores are uniform across the 256 dialogues, especially for clarity, where the dispersion is minimal (SD = 0.05). Naturalness (4.45 $\pm$ 0.09) was also highly rated, indicating that dialogues generally sounded realistic and conversational, though with slightly more variation compared to fluency. The relatively low entropy values across all dimensions reflect stable and confident model judgments. Structure received the lowest score (4.14 $\pm$ 0.12), pointing to occasional weaknesses in how model responses were organized or sequenced, which may offer room for future refinement in prompt design or dialogue planning.

Dimension	Agent Starter	Patient Starter
Fluency ($\pm$ SD)	4.70 $\pm$ 0.14	4.70 $\pm$ 0.14
Naturalness ($\pm$ SD)	4.46 $\pm$ 0.10	4.45 $\pm$ 0.09
Clarity ($\pm$ SD)	4.46 $\pm$ 0.05	4.48 $\pm$ 0.04
Structure ($\pm$ SD)	4.13 $\pm$ 0.13	4.15 $\pm$ 0.09

Table 4: Evaluation scores (mean $\pm$ SD) by dialogue starter type.

Looking at dialogue initiation at Table 4, we found no substantial difference in evaluation scores between agent- and patient-starting dialogues. Both groups scored almost identically across fluency, naturalness, and clarity. Patient-initiated dialogues had slightly higher clarity and structure, though the differences are marginal (≤ 0.02). Entropy values were also nearly identical, suggesting GPT-4o’s confidence did not vary depending on the first speaker. This indicates that our prompting setup yielded consistent dialogue quality regardless of turn-taking structure at the opening.

⁴OpenAI gpt-4o (August 2025 release)

A methodological consideration of this evaluation setup is the potential for self-preference bias: since the dialogues were generated by GPT-4 and evaluated by GPT-4o (a model from the same family), the evaluator may exhibit systematic leniency toward outputs that conform to its own generation patterns (Panickssery et al., 2024). To mitigate the impact of this concern, we complement the LLM-based evaluation with an independent human expert evaluation (4.3.2), which may capture quality dimensions that the LLM-based evaluation alone cannot.

4.3.2. Human-based Evaluation

To complement the LLM-based analysis, we conducted a human study using our Synthetic Patient-Agent Conversation Evaluation Questionnaire (SPACEQ), inspired by the Chatbot Usability Questionnaire (CUQ) (Holmes et al., 2019). The full questionnaire is available in the accompanying repository¹. Although the LLM-based evaluation was conducted over all 256 generated conversations, for the human evaluation, we selected 20 dialogues (covering all 16 persona profiles from Table 1) to limit cognitive load while ensuring representativeness.

Ten domain experts—researchers working on diabetes lifestyle support mostly affiliated with CHIP or related projects in the medical field—were instructed to read each dialogue carefully and rate it on a 5-point Likert scale (Likert, 1932) (1 = Very Poor, 2 = Poor, 3 = Acceptable, 4 = Good, 5 = Excellent), with optional free-text remarks after each item to capture explanations and qualitative feedback. The questionnaire content and rating rubric mirror the language-level criteria used in the LLM-based evaluation in Section 4.3.1, enabling a direct, dimension-wise comparison between human judgments and model-based scores.

Dimension	Avg. Human Score ($\pm$ SD)	α
Fluency	4.65 $\pm$ 0.54	−0.035
Naturalness	3.95 $\pm$ 0.89	0.020
Clarity	4.48 $\pm$ 0.77	0.021
Structure	4.54 $\pm$ 0.66	0.050

Table 5: Human evaluation over 20 dialogues and 10 expert raters. For each dimension, we report the mean and standard deviation (SD) across all ratings, alongside inter-rater agreement using Krippendorff’s α (ordinal).

As shown in Table 5, raters judged dialogues as highly fluent (4.65 $\pm$ 0.54) and clear (4.48 $\pm$ 0.77), but naturalness received the lowest score (3.95 $\pm$ 0.89). Higher standard deviations for naturalness and clarity suggest greater variability, with some dialogues considered realistic while others were described as “too rigid” or “repetitive”. In contrast,

fluency showed lower variability, indicating stronger agreement on grammatical correctness.

Inter-rater agreement was estimated with Krippendorff’s α (ordinal) (Krippendorff, 2011) across the 10 raters for the 20 conversations. The coefficients were close to zero, which at face value would suggest poor agreement. However, this result should be interpreted in the context of a well-documented statistical artifact: when ratings are concentrated within a narrow range, reliability coefficients such as Krippendorff’s α lose discriminative power due to restricted variance (Gwet, 2014). In our case, the vast majority of ratings fell between 4 and 5 on the 5-point scale, meaning that raters broadly converged in judging dialogues as “Good” or “Excellent”. The low α values therefore do not indicate disagreement but rather reflect insufficient rating dispersion for the coefficient to detect the convergence that is evident in the raw score distributions. Future evaluations could consider using a finer-grained scale (e.g., 7- or 10-point) or percent agreement metrics to better capture agreement under ceiling conditions.

Rater comments provide additional context for these results. For fluency, many notes recognized positively the language as “smooth and grammatically correct” and the explanations as “easy to follow”. On the other hand, for naturalness, several dialogues were described as templated or slightly rigid. Specifically, raters described parts of the dialogue as “robotic”, “scripted” or “less natural” due to frequent name repetition, abrupt topic shifts, or judgmental phrasing in single turns (e.g., “Frequent name repetition makes it sound less natural.”). For clarity and structure, participants consistently appreciated stepwise guidance and succinct, practical advice, with positive remarks highlighting clear sequencing such as “first... then... next...” and appropriate escalation when needed, as in “suggesting an appointment with a specialist when symptoms persist”. A few comments also pointed to tone, endorsing supportive phrasing in many cases while suggesting that some questions could be softened. Overall, the human evaluation confirmed strong linguistic quality, but identified naturalness as the main area where improvement is needed.

5. Triple Extraction

Following conversation creation, we extracted Subject–Predicate–Object (SPO) triples⁵ to provide structured representations for enriching CHIP’s Domain Knowledge Graph.

⁵Here Subject–Predicate–Object do not represent syntactic dependencies but rather semantic relations in the Semantic Web paradigm.

5.1. Conversation Annotation

We frame conversational triple extraction (CTE) as token-level sequence labeling followed by triple assembly. Each token is annotated as *Subject*, *Predicate*, *Object*, or *other*, allowing multiple tokens per role and tokens participating in multiple triples.

Given medical content sensitivity, we used manual annotation for validation (882 sentences) and test sets (996 sentences), with automated methods for the training set (2953 sentences). The schema extracts semantically meaningful SPO triples from patient utterances, while agent responses were excluded as their content is predefined in the system’s knowledge base. The annotation schema was developed to be robust enough to handle special cases, including passive voice, compound Subjects and Predicates, subordinate clauses, and implied elements. In addition to that, special rules were applied to question-answer sequences to avoid redundant labeling. For example, if an answer was merely affirmative, only the question was annotated; if the answer contained substantial information, only the answer was annotated; if both contributed relevant content, both were labeled. The full annotation schema can be found in the accompanying repository ¹.

Manual annotation was conducted using the INCEpTION tool (Klie et al., 2018) with three linguistic experts independently annotating a subset of the data. Fleiss’ Kappa of 0.657 indicated substantial agreement (Landis, 1977), with disagreements resolved through majority voting and expert review.

For the training set, annotation was performed automatically using structured prompting with GPT-4o. We reused the annotation prompt from the manual phase and applied a dynamic chunking strategy to handle question-answer pairs appropriately. The model was run at a low `temperature` (0.2) to ensure consistency and `max_tokens` of 4096 (default) to avoid truncation in longer dialogues. This approach allowed us to scale the annotation process efficiently while maintaining alignment with the manually labeled schema. The resulting corpus includes 4831 annotated sentences in total.

5.2. Methods

We evaluated three complementary approaches spanning symbolic baselines to state-of-the-art neural models, examining trade-offs between interpretable rules, general-purpose LLM prompting, and domain-adapted training. *Rule-based syntactic parsing* uses a dependency parser to identify predicates and attach Subjects/Objects, offering interpretability but sensitivity to conversational variation. *GPT-4o* applies few-shot demonstrations at `temperature` 0.2 without fine-tuning to test generalization. *Fine-tuned BERT* adapts `bert-base-`

System	Token-level (macro)			Triple matching (F1)				
	P	R	F1	Full	Partial	Subj	Pred	Obj
Rule-based	0.62	0.50	0.55	0.00	0.29	0.29	0.20	0.00
GPT-4o	0.72	0.65	0.68	0.35	0.64	0.58	0.43	0.43
BERT	0.73	0.63	0.67	0.26	0.50	0.47	0.39	0.32

Table 6: Token-level SPO classification and triple matching results.

`uncased` (Devlin et al., 2019) with a linear token classification head and Optuna-tuned hyperparameters, achieving validation macro-F1 of 0.7563 at learning rate 2.209×10^{-5} and batch size 1.

5.3. Evaluation Metrics

We report macro-averaged precision, recall, and F1 at the token level, plus triple matching metrics: *full* (all components match), *partial* (two components match), and *per-component matches* (one component match), accounting for informative partial triples.

6. Discussion

Our evaluation reveals complementary strengths when comparing LLM-based and human assessments. Both consistently rated fluency and clarity as strong, reflecting grammatically correct and understandable dialogues. However, differences emerged for naturalness and structure. GPT-4o assigned higher naturalness scores (4.45 vs. 3.95), while human raters flagged repetitive phrasing, rigid tone, and name overuse as reducing realism. Conversely, humans valued structure more (4.54 vs. 4.14), appreciating stepwise guidance and escalation. Although inter-rater agreement coefficients were low, this largely reflects ceiling effects: most ratings clustered between 4 and 5, limiting variability while still showing broad convergence that the dialogues were judged as good or excellent.

Triple extraction experiments provide further insights. As shown in Table 6, the rule-based approach while interpretable, struggled with conversational variation and produced very low *full match* scores (0.00). Both neural methods substantially outperformed this baseline. GPT-4o achieved the best token-level (F1 = 0.68) and triple-level results, with strong *partial match* performance (0.64), capturing meaningful fragments even when full triples were incomplete. Fine-tuned BERT performed slightly below GPT-4o overall (token-level F1 = 0.67; *partial match* = 0.50) but showed robustness to domain-specific patterns, demonstrating that domain adaptation benefits annotation style capture. A common challenge across models was low *full match* accuracy (GPT-4o = 0.35; BERT = 0.26), underscoring the difficulty of aligning all three triple

components consistently in conversational data. When examining single-component matches, GPT-4o achieved the strongest results across Subject (0.58), Predicate (0.43), and Object (0.43), while BERT followed closely, suggesting that both models can reliably identify individual components even when full triple alignment is incomplete.

These results demonstrate that our synthetic dataset is linguistically validated and effective for benchmarking triple extraction, while highlighting the need for hybrid strategies: LLM prompting offers broad generalization, while domain-adapted models capture contextual nuances more reliably.

7. Conclusion

This paper presented a methodology for generating and annotating a synthetic conversational dataset for Type 2 Diabetes management. Using demographically and medically grounded personas, we produced 256 multi-turn dialogues incorporating personalization, empathy, contextual awareness, and medically accurate advice. The dataset was enriched with Subject–Predicate–Object (SPO) annotations through manual expert labeling and automated LLM support, yielding 4831 annotated sentences.

Our evaluation combined LLM-based scoring and human expert judgments, confirming high linguistic quality while identifying naturalness as the main area for improvement. We benchmarked three triple extraction approaches—rule-based syntactic parsing, GPT-4o, and fine-tuned BERT with a linear token classification head—with GPT-4o achieving the strongest performance.

The dataset, annotation schema, and baseline results provide a new resource for developing and evaluating the knowledge extraction component of Hybrid Intelligence systems such as CHIP, and a reproducible benchmark for conversational triple extraction in patient-centered agents.

8. Limitations

Our findings highlight important limitations of the dataset itself. Patient turns tend to be well-formed and elaborate, lacking the spelling errors, informal phrasing, repairs, and fragmented utterances typical of real conversations. This may inflate fluency and clarity scores while masking difficulties in handling noisy input. Systems trained on this data may therefore need additional adaptation to handle the misalignments and ungrammatical constructions that characterize authentic patient communication.

Moreover, although CHIP is intended for patients in the Netherlands, the dataset was generated in English rather than Dutch due to resource availability. This language choice facilitates evaluation

with widely available models but limits immediate applicability to Dutch-speaking populations, where additional adaptation and validation would be required.

Finally, the dataset has been evaluated intrinsically—through linguistic quality assessments and triple extraction benchmarking—but has not yet been tested as training data within a deployed dialogue system. A natural next step would be to evaluate whether models trained on this data improve the accuracy of knowledge graph updates when deployed within the full CHIP pipeline during real or Wizard-of-Oz patient–agent interactions. This falls outside the scope of this paper but remains an important direction for future validation.

9. Ethics Statement

All dialogues used in this study were synthetically generated using large language models (LLMs) and do not contain any real patient information. No personally identifiable data were used or collected at any stage. The human evaluation involved voluntary participation of domain experts who provided consent for their anonymized responses to be analyzed in aggregate form. The study complies with the ethical principles of the participating institutions and adheres to the data protection requirements of the General Data Protection Regulation (GDPR).

10. Bibliographical References

- Jessica C. G. Bak, Dick Mul, Erik H. Serné, Harold W. de Valk, Theo C. J. Sas, Petronella H. Geelhoed-Duijvestijn, Mark H. H. Kramer, Max Nieuwdorp, and Carianne L. Verheugt. 2021. [Dpard: rationale, design and initial results from the dutch national diabetes registry](#). *BMC Endocrine Disorders*, 21(1):122.
- Centraal Bureau voor de Statistiek. 2024. [Gezondheid en zorggebruik; persoonskenmerken](#). CBS StatLine. Accessed: 2024-06-02.
- Maryam Changizi and Mohammad H. Kaveh. 2017. [Effectiveness of the mhealth technology in improvement of healthy behaviors in an elderly population—a systematic review](#). *mHealth*, 3(11).
- Anshul Chavda and Pushpak Bhattacharyya. Synthetic dialogue data generation: A comprehensive survey of methods, evaluation, and challenges.
- Floris den Hengst, Shaad Alaka, and Bart A. Kamphorst. 2026. [Collaborative hybrid intelligence](#)

- platform chip: A modular architecture for developing and testing personalized lifestyle support interactions. *SoftwareX*, 33:102536.
- 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.
- Bernd JW Dudzik, Jasper S van der Waa, Pei-Yu Chen, Roel Dobbe, Ínigo MDR de Troya, Roos M Bakker, Maaïke HT de Boer, Quirine TS Smit, Davide Dell'Anna, Emre Erdogan, et al. 2024. Hybrid intelligence supports application development for diabetes lifestyle management. *Journal of Artificial Intelligence Research*, 80:919–929.
- Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.
- Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *Proceedings of the 31st European conference on cognitive ergonomics*, pages 207–214.
- Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*.
- Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. [The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation](#). In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.
- Klaus Krippendorff. 2011. [Computing krippendorff's alpha-reliability](#).
- JR Landis. 1977. The measurement of observer agreement for categorical data. *Biometrics*.
- Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.
- Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.
- 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](#).
- OpenAI. 2024. [Gpt-4 technical report](#).
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. [Llm evaluators recognize and favor their own generations](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802. Curran Associates, Inc.
- Tessa F Peerbolte, Rozanne JA van Diggelen, Pieter van den Haak, Kim Geurts, Luc JW Evers, Bastiaan R Bloem, Nienke M de Vries, and Sanne W van den Berg. 2025. [Conversational agents supporting self-management in people with a chronic disease: Systematic review](#). *J Med Internet Res*, 27:e72309.
- Krithika Ramesh, Nupoor Gandhi, Pulkit Madaan, Lisa Bauer, Charith Peris, and Anjalie Field. 2024. Evaluating differentially private synthetic data generation in high-stakes domains. *arXiv preprint arXiv:2410.08327*.
- Selene Báez Santamaria, Lea Krause, Lucia Donatelli, and Piek Vossen. 2023. The role of personal perspectives in open-domain dialogue.
- Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2023. Data augmentation for conversational ai. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5220–5223.
- Joanne K. Ujcic-Voortman, Miranda T. Schram, Monique A. Jacobs-van der Bruggen, Arnoud P. Verhoeff, and Caroline A. Baan. 2009. [Diabetes prevalence and risk factors among ethnic minorities](#). *European Journal of Public Health*, 19(5):511–515.
- E H Wagner, B T Austin, and M Von Korff. 1996. [Organizing care for patients with chronic illness](#). *The Milbank Quarterly*, 74(4):511–544.

11. Language Resource References

- Steven Bedrick and A. Seza Doğruöz and Sergiu Nisioi. 2025. [A Typology of Synthetic Datasets for Dialogue Processing in Clinical Contexts](#).

- Trisha Das and Dina Albassam and Jimeng Sun. 2024. *Synthetic Patient-Physician Dialogue Generation from Clinical Notes Using LLM*.
- Demszky, Dorottya and Movshovitz-Attias, Dana and Ko, Jeongwoo and Cowen, Alan and Nemade, Gaurav and Ravi, Sujith. 2020. *GoEmotions: A dataset of fine-grained emotions*.
- Eisenberg, Joshua and Sheriff, Michael. 2020. *Automatic extraction of personal events from dialogue*.
- Ahmad Rezaie Mianroodi and Amirali Rezaie and Niko Grisel Todorov and Cyril Rakovski and Frank Rudzicz. 2025. *MedSynth: Realistic, Synthetic Medical Dialogue-Note Pairs*.
- Wang, Junda and Yao, Zonghai and Yang, Zhichao and Zhou, Huixue and Li, Rumeng and Wang, Xun and Xu, Yucheng and Yu, Hong. 2024. *NoteChat: A Dataset of Synthetic Patient-Physician Conversations Conditioned on Clinical Notes*. Association for Computational Linguistics.
- Yu, Dian and Yu, Zhou. 2019. *Midas: A dialog act annotation scheme for open domain human machine spoken conversations*.

Italian Medical Term Simplification: From Patient Information Leaflets to Simplified Language Resources

Maria Pia di Buono

University of Naples "L'Orientale"
Via Duomo, 219 - 80136 - Napoli (Italy)
mpdibuono@unior.it

Abstract

Terminological simplification in patient information leaflets (PILs) is implemented through a variety of linguistic strategies. Although these strategies help improve text comprehensibility, their overall impact remains limited. The complexity of PILs is still influenced by multiple factors, including frequent cross-referential terminology, the presence of subordinate clauses, lengthy sentences, and the use of domain-specific terms. This paper introduces I-MTS (Italian Medical Term Simplification), the first resource specifically developed for medical term simplification in Italian. I-MTS is designed to support research on lexical simplification and to facilitate the automatic adaptation of medical texts for non-expert audiences, thereby enhancing the readability and accessibility of health information in Italian.

Keywords: Medical Term Simplification, Italian Linguistic Resources, Health Literacy

1. Introduction

Promoting clear communication and facilitating access to medical information are central objectives in healthcare, as reflected in the Guideline on the Readability of the Labelling and Package Leaflet of Medicinal Products for Human Use issued by the European Medicines Agency (EMA).¹

In line with this, also the International Plain Language Federation (IPLF) has developed guidance to enhance healthcare communication through the application of the ISO Plain Language Standard to health information.² Drawing on the widely adopted definition of plain language, the guide outlines a set of core principles designed to ensure that health information is relevant, findable, clear, and usable by patients and the public.

These efforts align with the broader aim of promoting health literacy (Simonds, 1974), intended as "the degree to which individuals and communities have the capacity to obtain, process, understand, evaluate, and apply information needed to make public health decisions that benefit the community" (Freedman et al., 2009). Thus, health literacy - encompassing both individual and collective dimensions - may play a central role in empowering patients to make informed decisions and actively participate in their own healthcare.

As an attempt to apply such indications to promote health literacy and improve healthcare communication, and with the aim of mitigating the use of medical jargon in patient information leaflets (PILs), simplifications are incorporated into package insert texts through a variety of linguistic strategies.

Examples³ of these strategies include paraphrastic or explicative reformulations such as colestipol and colestiramina, described as *medicines used to reduce the intestinal absorption of fats*, as well as the use of clarifying parenthetical expressions such as *sexual desire (libido)*.

Although these strategies contribute to improving text comprehensibility, their overall effect remains limited, as several factors continue to contribute to the complexity of PILs. Among these are the presence of subordinate clauses and lengthy sentences even in simplified texts, the inconsistency of term simplification throughout the text, the frequent use of cross-referential terminology - for example, *medicines belonging to the same group as Akineton ampoules* (anticholinergic drugs) - as well as the coexistence of domain-specific terminology.

Indeed, simplification in PILs tends to operate locally, targeting specific lexical items or short appositive phrases, while the overall lexical and syntactic complexity of the text often remains unchanged. Expressions such as quinolone antibiotics, *used to treat infections* illustrate how specialized terminology persists and how simplifications may rely on non-nominal constructions that can still challenge readers' comprehension, even within otherwise simplified contexts.

Thus, despite initiatives to promote health literacy and improve healthcare communication, PILs often remain overly complex, hindering patients' full understanding and underscoring the need for systematic approaches to identify and model simplification strategies at both the lexical and syntactic levels.

³In the examples, technical terms are underlined, and simplified expressions are shown in italics. When present, Italian examples are marked by quotation marks, and English translations are provided in parentheses.

To address these challenges, we introduce I-MTS (Italian Medical Term Simplification), the first resource specifically developed for the simplification of medical terminology in Italian. I-MTS has been designed to support research on lexical simplification and to facilitate the automatic adaptation of medical texts for non-expert audiences, thereby enhancing the readability and accessibility of health information in Italian.

The paper is structured as follows: Section 2 reviews related work in the field, Section 3 details the methodology employed at each stage, Section 4 presents an overview of the resulting linguistic resource, and Section 5 concludes with a discussion of the findings and their limitations.

2. Related Work

Several studies have addressed the problem of medical text simplification, exploring a range of topics and languages, from the development of simplified linguistic resources to the evaluation of simplification strategies and user comprehension. Among these, we highlight only a selection of the most relevant contributions.

The development of linguistic resources address both the identification of complex terminology and the creation of parallel data for training and evaluation to support medical text simplification. In this direction, some approaches aim to enhance interpretability by generating explicit explanations of specialized terms, for instance by leveraging semantic type information and predefined connectors to produce structured paraphrases (Kandula et al., 2010). Other work concentrates on the systematic identification of medical jargon in real-world texts, combining Natural Language Processing (NLP) models with the construction of expert-annotated datasets, such as MedJ, a corpus of over 18,000 sentences from electronic health record (EHR) notes designed to support research on jargon extraction and readability enhancement (Kwon et al., 2022). More recent efforts have moved towards the development of parallel corpora specifically tailored to simplification tasks, including resources composed of manually simplified educational materials, as Simplified Digestive Cancer (SimpleDC) (Rahman et al., 2024), to enable the investigation of advanced simplification approaches based on large language models, including fine-tuning, reinforcement learning, and prompt-based methods. Alongside the development of linguistic resources, a growing body of work has also addressed the evaluation of medical text simplification. In this context, Devaraj et al. (2021) introduce a parallel corpus of technical and lay summaries in English, covering multiple clinical topics, and propose a novel evaluation metric based on likelihood scores derived

from masked language models pretrained on scientific texts, thus providing an automatic approach to assessing simplification quality.

Complementary approaches have investigated the assessment and improvement of lexical complexity in medical texts, with a particular focus on terminology. For example, Abrahamsson et al. (2014) adapt a method for estimating word difficulty in the Swedish medical domain by combining frequency-based measures with more fine-grained features, such as substring frequency, in order to account for language-specific phenomena like compounding to identify and replace complex terms.

Similarly, other studies have explored the potential of existing consumer-oriented lexical resources, evaluating both their coverage and their suitability for simplifying specialized texts. In this line of work, Qenam et al. (2017) examine the comprehensibility and suitability of the Consumer Health Vocabulary (CHV) in terms of its ability to support the simplification of radiology reports, through analyses of lexical similarity and content alignment between technical terms and their preferred consumer equivalents. Finally, Bala et al. (2020) investigate patient comprehension and perceptions of AI-generated simplified medical notes. Through comprehension questions and guided interviews, they assess perceived benefits and concerns. This research is closely related to the OpenNotes⁴ initiative, which began as an effort to share medical notes with patients to enhance understanding and engagement. While OpenNotes has shown benefits in improving patient satisfaction and clinical outcomes, clinician concerns about potential misunderstandings and increased workload remain.

3. Methodology

Despite the limitations that PILs still present, they constitute a valuable source of term simplifications that can be leveraged to identify and extract simplified variants of technical terms for the development of a domain-specific linguistic resource.

In line with this perspective, rather than generating or manually validating simplifications, the present study adopts a data-driven methodology to extract real-world term-simplification (T-S) pairs as they naturally occur in the context of PILs.

For this purpose, we select authoritative data released under an open license to model intralinguistic equivalence between specialized terms and their simplified counterparts, and to develop a structured resource in which each T-S pair is enriched with additional metadata, such as variant forms and semantic classes.

The following subsections describe each step of the methodology in detail.

⁴<https://www.opennotes.org/>

3.1. Data Selection

AIFA⁵, the Italian Medicines Agency, follows Open Data principles and distributes its resources under a CC-BY 4.0 license. Its datasets include pharmaceutical and "transparency lists"⁶, which provide comprehensive electronic inventories of drugs with metadata such as active ingredient, packaging, and Marketing Authorisation (MA) codes. AIFA also maintains a public database⁷ containing all summaries of product characteristics (SmPCs) and PILs approved by AIFA or EMA.

Using the AIFA resources, [Giordano and Di Buono \(2024\)](#) created the D-LeafIT corpus, formed by 1,819 Italian drug leaflets, of which 1,439 are for generic drugs and 380 for Class A drugs. The D-LeafIT corpus has served to evaluate the capability of large language models to provide reliable drug information to patients.

Subsequently, [Manna et al. \(2024\)](#) extracted a random subset of 600 entries to develop the Drug Information Mining in Italian (DIMMI) dataset as part of CALAMITA ([Attanasio et al., 2024](#); [Nissim et al., 2025](#)), a shared task - sourced by a community effort - to test the LM abilities in Italian, encompassing several different type of tasks.

Besides, the commercial drug names and their correspondent PILs raw text, DIMMI contains a manually annotated gold standard (GS) designed for the shared-task evaluation. For each of the drug entry, the DIMMI GS provides a list of entities present in the related PIL, covering five classes: i) molecule; ii) dosage; iii) drug interaction; iv) usage; and v) side effects.

The DIMMI GS annotation was carried out by three annotators, following detailed guidelines that primarily addressed span detection rather than category assignment, the latter being relatively straightforward. The final inter-annotator agreement (IAA) reported in the paper indicates a good level of agreement across all categories.

In particular, we focus on three of the five categories annotated for each entry in DIMMI GS: drug interaction, side effects, and usage. This choice is motivated by two considerations. First, as reported by the authors, these categories achieved the highest IAA scores in the annotation, measured as average pairwise F1 for usage (0.68) and token-level agreement percentage for drug interaction (0.66) and side effect (0.78). Second, a randomized manual inspection of the data indicates that these three categories exhibit the highest frequency of terminological simplifications in PILs. The total number of DIMMI GS not-unique entities for the three selected categories are 49,016 (Table 1).

⁵AIFA - Homepage

⁶AIFA - Pharmaceutical and Transparency Lists

⁷AIFA - Banca Dati Farmaci

3.2. Intralinguistic Equivalence Modelling

Specifically designed to evaluate the ability of language models to answer targeted questions about drugs in Italian extracting information from PILs, the DIMMI GS annotation does not provide an explicit distinction between technical terms and their simplified variants.

Furthermore, as information in PILs are presented with an uneven level of details, different annotation choices regarding the levels of conceptual granularity across categories have been adopted in DIMMI ([Manna et al., 2024](#)).

In general, the minimal-span principle and the most-specific-concept principle are applied. For example, in "meningite criptococcica (*un'infezione micotica del cervello*)" (*cryptococcal meningitis (a fungal infection of the brain)*), only the most specific term, meningite criptococcica, is annotated.

In other cases, when only generic conceptual entities are presented, e.g., "*antibiotici*" (*antibiotics*), the information-preservation principle is applied.

As a result, the list of entities in the DIMMI GS may include both specialized terms and their simplified counterparts, annotated as two distinct entities. This implies that the intralinguistic equivalence between a term and its corresponding simplification, when present in a PIL, is not fully preserved.

To restore the equivalence between specialized terms and their simplified counterparts, we adopt a methodology similar to the one proposed by [Speranza et al. \(2021\)](#), which relies on the use of appositives to extract term simplifications in the Cultural Heritage domain.

Due to their semi-fixed syntactic patterns and high degree of semantic transparency, appositive constructions can be used as reliable linguistic indicators of simplified language ([Speranza et al., 2021](#)). Syntactic constructions of appositives, usually marked with the presence of brackets, commas, and hyphens, can be systematically described and exploited to identify the term and its corresponding simplified reference.

Although the apposition-based extraction method could, in principle, be applied to raw texts independently of a predefined list of entities, we deliberately anchor the process to the entities already annotated in the DIMMI GS. This choice allows us to (i) preserve and fully exploit the manual annotations available in DIMMI; (ii) reduce the potential noise in the apposition-based extraction of simplifications; (iii) ensure methodological consistency and comparability with the existing resource.

Apposition Extraction Before proceeding with apposition extraction, we examined the target entities in the DIMMI GS to identify entries that already contained parenthetical content. These cases re-

flect annotation variability, specifically instances in which entities were not annotated according to the minimal-span and most-specific-concept guidelines. Such variability can be attributed to the aforementioned challenges in span definition, which are also reflected in the IAA scores. We found that the annotated entities presenting a parenthetical content are about 19,90% of the total annotated entities, with a variability among categories (see Table 1). These entities undergo a cleaning procedure aimed at identifying and removing potential annotation inconsistencies. We evaluate them according to two criteria: (i) the entry must not begin with a punctuation mark; and (ii) its syntactic structure must correspond to that of a well-formed noun phrase in Italian. The former is implemented through regular expressions designed to detect leading punctuation characters and ten entries are removed. The latter is assessed via automatic part-of-speech (POS) tagging, by verifying that the internal structure of each span conforms to a well-formed Italian noun phrase pattern. Entries beginning with a preposition - an invalid structure for Italian noun phrases - are excluded (157 entities).

Violations of these criteria are interpreted as evidence of incomplete span annotation, suggesting that relevant elements of the entity are missing. Since such cases may compromise the reliability and consistency of subsequent analyses, these entities are excluded from further processing. The total number of retained entities is presented in Table 1.

For DIMMI GS entities without an annotated paren-

Category	+ Par. [Ret.]	- Par.	Total
Drug Int.	979 [834]	9,690	10,669
Side Eff.	8,268 [7,149]	27,985	36,253
Usage	508 [414]	1,586	2,094
Total	9,755 [8,397]	39,261	49,016

Table 1: Number of DIMMI entities with (+ Par.) parenthetical content total and retained [Ret.] and without (- Par.) parenthetical content per each category

thetical content, concordance searches were conducted in the DIMMI PIL corpus using a context window that dynamically adapted to each occurrence. The inclusion of an entity and its surrounding content was guided by the presence or absence of parentheses, ensuring that the context captured was appropriate for the structure of each instance. In the absence of specific guidelines governing the simplification process, intralinguistic equivalence in PILs may be realized through appositive structures that vary in the order of their components and/or in the markers used to signal simplification. For instance, an equivalence between the term "acufene" (tinnitus) and the simplification "ronzio

alle orecchie" (ringing in the ears) can be realized differently within different texts, as in the examples below.

"ronzio alle orecchie (acufene)"

"acufene (ronzio alle orecchie)"

To account for such structural variation in appositive constructions, we employ regular expressions to model the possible combinatorial patterns involving the targeted DIMMI entities.

Since not all PILs provide simplifications for every term, some of the entities in the selected class from the DIMMI GS may not belong to a T-S pair within a given text. To exclude these standalone entities from the analysis, entries that did not display markers of appositive constructions in their context were removed.

By verifying the presence of appositive structures, we were able to identify entities that qualify as candidates for inclusion in a T-S pair and to extract them, together with their appositive counterparts, from the unstructured texts in which they occur.

The total numbers of T-S pairs resulting from this step is presented in Table 2.

Pair Filtering The extracted appositive components do not always instantiate an intralinguistic equivalence by providing a simplified version of the target term. Instead, they may realize other types of semantic relations.

For example, in "traumi (ad esempio causati dallo sport)" [injuries (e.g., sports-related)], the parenthetical content has an illustrative function and does not introduce a semantic equivalence through a distinct, reusable concept. As such cases do not qualify as T-S pairs, candidate entities have been filtered on the basis of the semantic relations they encode.

Given that medical language is highly specialized and is characterized by recurrent structural patterns and domain-specific lexical usage, we define some selection criteria to filter out appositive elements that do not correspond to term simplifications from the list of candidate pairs.

Syntactic and lexical constraints are applied to identify the semantic relation existing between the term and the second component in the candidate pairs. In particular, we exclude appositive-construction pairs in which the parenthetical segment encodes relations other than intralinguistic equivalence. These cases include:

- Coordinated structures introduced by simple coordinating conjunctions: copulative ("e", and, "né", nor), expressing additive coordination; adversative ("ma", but), expressing contrast; disjunctive ("o", or), expressing alternation or exclusion; as well as complex conjunctions (e.g., "anche se", even if).

- Non-complement subordinate clauses (i.e., adverbial or peripheral clauses), which function as clause-level adjuncts and encode semantic relations external to the argument structure of the main clause. These include: (a) temporal ("quando / mentre / finché / dopo che / prima che"); (b) causal ("perché / poiché / siccome"); (c) purpose ("perché / affinché"); (d) concessive ("sebbene / benché"); (e) conditional ("se / qualora / purché"); (f) result/consecutive ("così... che / a tal punto che"); (g) comparative ("più/meno [...] che, tanto [...] quanto"); (h) limitative/exceptive ("a meno che / per quanto").
- Adverbial phrases introduced by exemplificatory markers such as "per esempio" or "ad esempio" (for example), which signal illustrative expansion rather than reformulation or equivalence.
- Prepositional phrases functioning as clausal- or event-level adjuncts. However, when such phrases contribute to intra-domain explicitation rather than circumstantial modification - by encoding conceptual specification, referential restriction, or therapeutic indication (e.g., anatomical specification or functional purpose) - they are retained. These phrases are typically signalled in Italian through the use of "per" and articulated forms of "di" (e.g., "del / della / dello").

We retain prepositional phrases when they instantiate domain-conventionalized relational patterns between ontological classes (e.g., drug interaction, usage). In such cases, the phrase encodes concept-level relational specification (e.g., therapeutic indication, anatomical localization) rather than event-level circumstantial modification. The inclusion criterion is therefore based on the semantic class of the head noun and the ontological type of the complement, rather than on the preposition *per se*.

The criteria emerged from iterative corpus analysis and were subsequently operationalized as rule-based filters, combining data-driven generalizations with independently established morphosyntactic constraints of Italian. The number of pairs retained after the filtering phase is presented in Table 2.

Term-Simplification Alignment With the exception of retained prepositional phrases, which transparently encode simplification directionality, structural variation in appositive constructions prevents automatic identification of term *vs.* simplification within T-S pairs. Intralinguistic equivalence does not by itself determine role assignment. A dedicated reconciliation step is therefore introduced to

align pair components and resolve directionality. To differentiate entities representing technical terms from those representing simplifications, ensuring terminological validity and avoiding arbitrary term selection, T-S pair component were cross-checked against an external authoritative clinical reference (the Italian MSD Manual - Professional version⁸). The MSD Manual provides several resources including tables that report information about diseases and other topics. The MSD section headings and table titles were consulted for coverage verification⁹. These labels were subsequently normalized, lemmatized, and stripped of descriptive modifiers before comparison with our data. The resulting list constitutes our internal lexical benchmark suitable to identify the technical term within each T-S pair. A total of 1005 headings were examined, of which 589 corresponded to identifiable disease-related terms after normalization.

To account for the other entries in our T-S pair list we use the Wikipedia list for the "Farmaci secondo la funzione" (medication classified for function) category¹⁰. The category tree, comprising 162 unique terms, was retrieved and subsequently applied to verify our entries. To identify further terms related to anatomy concepts, we relied on an external encyclopedic list available online, which is one of the few such resources and originates from a private hospital website¹¹.

The pair items that we were not able to check against an authoritative source have been dismissed.

The number of pairs retained after the T-S alignment phase is presented in Table 2.

Data	Drug_Int.	Side_Eff.	Usage	Tot.
DIMMI	10,669	36,253	2,094	49,016
AE	3,122	9,851	691	13,609
PF	1,952	5,275	477	7,704
TSA	960	2,502	263	3,726

Table 2: Number of entities for the selected categories presented for each step of the intra-linguistic equivalence modelling: Apposition Extraction (AE); Pair Filtering (PF); Term-Simplification Alignment (TSA)

⁸Manuale MSD - Versione per i professionisti

⁹For the sake of methodological transparency, no expressive or structural elements of the source were reused, nor was any portion of the original database reproduced or redistributed. The resulting resource relies exclusively on independently curated data, enriched with original metadata and classification criteria.

¹⁰Wiki category list

¹¹Gavazzeni website

3.3. Final Reconciliation

After identifying the technical term within each T-S pair using an external resources, all entries were processed through a final reconciliation procedure to systematically reconcile morphological, orthographic, and simplification variations while preserving functional distinctions.

First, we investigated instances of partial simplifications. By partial simplification, we refer to cases in which the simplification is only partially realized, with the noun phrase head expected in the simplified form being absent. As a result, partial simplifications are typically realized as prepositional phrases (e.g., *(allo stomaco)*) or adjective phrases rather than full noun phrases. As final step of the final reconciliation, prepositional- and adjective-phrase simplifications were completed where possible, allowing for a consistent representation of simplifications while retaining distinctions critical to downstream analyses.

Then, inflection-aware consolidation was performed: singular and plural forms were initially retained to capture maximal surface form variability and to account for differences in the paired simplification component. Lemma-level grouping was then applied to reduce purely inflectional redundancy without collapsing semantically or functionally distinct TS mappings.

Next, a surface-level normalization step addressed orthographic and capitalization variants, ensuring harmonization of superficial differences.

A final grouping and cleaning procedures are performed to remove duplicates, irrelevant entries, and formatting inconsistencies.

This multi-step procedure ensured that morphological, orthographic, and simplification variations were systematically reconciled, while maintaining the structural and functional integrity of each T-S pair. The number of pairs retained after the final reconciliation phase is presented in Table 3.

3.4. Semantic Classification

Finally, semantic classification organizes the terms into predefined categories, facilitating downstream analysis and evaluation. This systematic post-processing ensures the quality, accuracy, and usability of the extracted data for subsequent tasks. DIMMI annotation labels are directly derived from the titles of PIL sections. To support a more general and reusable annotation scheme for NLP and language modeling in Italian, we map these labels onto a smaller set of more general semantic classes, using the classification derived from the term control lists. Specifically, most of the entries classified as drug interaction in DIMMI GS are retrieved from the Wiki control list, as this section predominantly

refers to interactions between drugs¹², and mapped to *medication*.

Similarly, entries in the usage class were retrieved from the disease control list and mapped to *disease*, as the corresponding section typically describes the medical conditions for which the drug is prescribed. Finally, we introduce the general class *body_part*, derived from the anatomy concept list. Unlike other mappings, this class is not tied to a single PIL section, as references to anatomical entities occur across different sections, most notably in side effects.

The introduction of this class reflects a shift from section-based labels to concept-based categories: entities denoting anatomical structures are grouped together regardless of their textual location, ensuring greater semantic consistency across the dataset.

4. I-MTS Overview

I-MTS (Italian Medical Term Simplification) is the first open-data resource specifically designed for the simplification of medical terminology in Italian¹³. It has been developed to support research on lexical simplification and to facilitate the automatic adaptation of medical texts for non-expert audiences, thereby improving the readability and accessibility of health information. The resource comprises 1356 T-S pairs in total¹⁴ (Table 3).

Category	# Entries
Disease	956
Medication	312
Body_Part	88
Total	1356

Table 3: Number of I-MTS pairs for Disease, Medication, Body Part Classes

Each entry in I-MTS is structured to include different types of lexical and semantic information (Table 4). For each entry, the resource includes: (i) a classification into three classes (CLASS) (i.e., Disease, Medication, and Body Part); (ii) available term variants (VAR), such as orthographic variations (e.g., "sindrome di Stevens Johnson" and "sindrome di Stevens-Johnson") or linguistic variants (e.g., "infarto del miocardio" and "infarto miocardico"), and linguistic borrowings or Latin-based forms, e.g., "miastenia gravis"; (iii) hierarchical subtypes (TYPE

¹²While drugs may also interact with nutraceuticals, such cases are relatively rare in the DIMMI GS.

¹³<https://github.com/banshee44/I-MTS>

¹⁴It is worth stressing that the total number of final entries exceeds that of the control lists. This reflects the fact that our resource captures multiple contextualized or variant forms of a concept, whereas the control lists provide a more limited, canonical representation.

CLASS	TERM	TYPE OF	VAR	SIMPL.
Disease	Ulcera	N/A	N/A	Lesione
Disease	Ulcera esofagea	Ulcera	N/A	Lesione della parete interna dell'esofago
Medication	Inibitori della cox-2	Inibitori	N/A	Medicinali che riducono l'infiammazione
Body Part	Uretra	N/A	N/A	Tubo che trasporta l'urina dalla vescica

Table 4: Examples of I-MTS entries

OF) are specified. These are identified through the partial overlap of multi-word terms with single terms, when this involves the head of the compound, e.g., "ulcera esofagea" and "ulcera boccale" are classified as subtypes of "ulcera".

Finally, all corresponding simplifications for each term are provided, resulting from the consolidation of the different variants attested across PILs, where the same term may be simplified in different ways.

5. Conclusions and Future Work

The current version of I-MTS represents an initial step toward the development of a richer and more informative resource for the simplification of medical terminology in Italian. Given the breadth of the medical domain, further efforts will be required to extend coverage to specific subdomains. Nevertheless, even in its present form, the resource can contribute to improving patients' health literacy.

At the same time, the resource remains preliminary and data-driven, as it is based on the systematic collection of existing simplifications rather than on validated or experimentally tested strategies. No quantitative measures of readability or comprehension have been computed, nor have user studies or formal validations against a gold standard been conducted. In addition, the dataset does not include measures of annotation consistency or inter-annotator agreement, and it does not aim to exhaustively capture all relevant linguistic phenomena. As such, it should be regarded as an exploratory and illustrative resource.

Despite these limitations, the dataset provides a useful foundation for future research. Although the present study does not include a systematic analysis of simplification strategies, the collected data can support the identification of recurring patterns, as well as the training, evaluation, and fine-tuning of language models for domain-specific text simplification. By aggregating and organizing real-world simplifications, I-MTS offers a data-driven basis for the development of more consistent and potentially consensus-driven approaches to medical text simplification.

Future work will focus on expanding the resource through the inclusion of additional patient information leaflets, formalizing annotation layers aligned with standardized classification systems to improve interoperability, and providing a systematic de-

scription of the simplification strategies employed. These efforts aim to enhance the resource's utility for both NLP applications and the development of consistent, evidence-based approaches to medical text simplification.

6. Acknowledgements

This research was carried out within the 4MLP Project, funded by the University of Naples "L'Orientale" under the University Research Projects funding scheme (PRA 2025) - CUP C66125002860005.

7. Bibliographical References

- Emil Abrahamsson, Timothy Forni, Maria Skeppstedt, and Maria Kvist. 2014. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 57–65.
- Giuseppe Attanasio, Pierpaolo Basile, Federico Borazio, Danilo Croce, Maria Francis, Jacopo Gili, Elio Musacchio, Malvina Nissim, Viviana Patti, Matteo Rinaldi, et al. 2024. Calamita: Challenge the abilities of language models in italian. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1054–1063.
- Sandeep Bala, Angela Keniston, and Marisha Burden. 2020. Patient perception of plain-language medical notes generated using artificial intelligence software: pilot mixed-methods study. *JMIR formative research*, 4(6):e16670.
- Ashwin Devaraj, Byron C Wallace, Iain J Marshall, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4972.
- Darcy A Freedman, Kimberly D Bess, Holly A Tucker, David L Boyd, Arleen M Tuchman, and Kenneth A Wallston. 2009. Public health literacy

- defined. *American journal of preventive medicine*, 36(5):446–451.
- Luca Giordano and Maria Pia Di Buono. 2024. Large language models as drug information providers for patients. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pages 54–63.
- Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. In *AMIA annual symposium proceedings*, volume 2010, page 366.
- Sunjae Kwon, Zonghai Yao, Harmon S Jordan, David A Levy, Brian Corner, and Hong Yu. 2022. Medjex: A medical jargon extraction model with wiki’s hyperlink span and contextualized masked language model score. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 11733.
- Raffaele Manna, Maria Pia Di Buono, and Luca Giordano. 2024. Dimmi-drug information mining in italian: A calamita challenge. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1144–1152.
- Malvina Nissim, Danilo Croce, Viviana Patti, Pierpaolo Basile, Giuseppe Attanasio, Elio Musacchio, Matteo Rinaldi, Federico Borazio, Maria Francis, Jacopo Gili, et al. 2025. Challenging the abilities of large language models in italian: a community initiative. *arXiv preprint arXiv:2512.04759*.
- Basel Qenam, Tae Youn Kim, Mark J Carroll, and Michael Hogarth. 2017. Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: translation and evaluation. *Journal of medical Internet research*, 19(12):e417.
- Md Mushfiquir Rahman, Mohammad Sabik Irbaz, Kai North, Michelle S Williams, Marcos Zampieri, and Kevin Lybarger. 2024. Health text simplification: An annotated corpus for digestive cancer education and novel strategies for reinforcement learning. *Journal of Biomedical Informatics*, 158:104727.
- Scott K Simonds. 1974. Health education as social policy. *Health education monographs*, 2(1_suppl):1–10.
- Giulia Speranza, Maria Pia Di Buono, and Johanna Monti. 2021. Terms and appositions: What unstructured texts tell us. In *International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ*, pages 219–230. Springer.

Evaluating Professional Acceptability of LLM-Generated Systematic Review Summaries in Healthcare: Psychiatrists' Perspectives

Paul Thompson¹, Artemis Boulogeorgou², Fotini Kaponi², Efstathia Soufleri²
and Sophia Ananiadou^{1,2}

¹National Centre for Text Mining, School of Computer Science, University of Manchester

²Archimedes/Athena Research Center, Greece

{paul.thompson, sophia.ananiadou}@manchester.ac.uk

bo.artemis@gmail.com, fotinikaponi@hotmail.gr, e.soufleri@athenarc.gr

Abstract

Cochrane systematic reviews evaluate the effectiveness and safety of medical interventions. Patients can benefit from clinicians' integration of outcomes of these reviews into their daily practices. However, systematic reviews are usually long documents; even their abstracts can extend to 1000 words, making rapid appraisal challenging for busy health professionals. Large language models (LLMs) offer potential to further distil these abstracts. Nevertheless, generating high-quality, clinician-oriented summaries in this context is non-trivial. They must comprehensively cover the original abstract, while remaining accurate and *professionally acceptable*, i.e., retaining all clinically important details. To address this challenge, we have developed a novel dataset, PsycSumEval, comprising summaries generated by four different LLMs for 115 Cochrane abstracts concerning mental health. Psychiatrists evaluated each summary across nine content dimensions, assigning scores and providing free-text justifications that highlight inaccuracies and missing details. The corpus provides fine-grained insight into how psychiatrists assess professional acceptability of compressed medical evidence. Rather than treating agreement as a merely statistical endpoint, we capture structured expert judgments alongside their rationales, enabling transparent analysis of where professional norms are stable and where interpretive latitude persists. We contribute both a rigorous evaluation dataset and an explicit model of expert acceptability criteria for medical evidence summarisation.

Keywords: Systematic Reviews, Summarisation, Evaluation Framework, Large Language Models

1. Introduction

Systematic reviews are the cornerstone of evidence-based medicine (Higgins et al., 2024; Antes et al., 2022; Dickersin et al., 1994; Chalmers and Altman, 1995; Needleman, 2002). By identifying, appraising and synthesising data from multiple trials and observational studies, such reviews provide reliable overviews of the advantages and disadvantages of medical interventions. These overviews can benefit patients, e.g., by allowing clinicians to integrate the latest research findings into their daily practices, or to support the development of clinical practice guidelines aimed at optimising patient care (Morton et al., 2011).

Since the complex nature of the review process results in long documents, busy health professionals may rely on the accompanying abstracts to acquire an understanding of the most important aspects of the review (Tsou and Treadwell, 2016). Past concerns about poor quality abstracts (Beller et al., 2011) were part of the motivation behind the development of the now widely-endorsed PRISMA guidelines (Liberati et al., 2009; Page et al., 2021a,b), which aim to ensure quality and consistency in systematic reviews of healthcare interventions. A dedicated checklist for abstracts (PRISMA-A) (Beller et al., 2013) aims to ensure

that the abstracts of such reviews can stand alone in presenting a clear and truthful account of the research. Particular journals or organisations may extend PRISMA-A with their own requirements, e.g., Cochrane systematic reviews must supplement results with assessments of evidence certainty (Antes et al., 2022).

The introduction of PRISMA-A has resulted in a steady increase in the median length of Cochrane abstracts, rising from 353 words in 2000 to 838 words in 2022 (Helbach et al., 2023b). However, evidence that guideline compliance can be achieved using far fewer words (Helbach et al., 2023a; Maticic et al., 2019) suggests that Cochrane abstracts could be further compressed, without sacrificing vital details. This would be especially beneficial for busy health professionals seeking rapid appraisal of multiple reviews.

Given the large amount of research into their use for biomedical text summarisation (Xie et al., 2023; Huang et al., 2025), large language models (LLMs) constitute an attractive option to produce summaries of Cochrane abstracts automatically. Indeed, previous work has investigated the LLM-driven generation of plain language summaries of these abstracts for non-expert readers (Devane et al., 2026; Arias-Russi et al., 2025a,b). Here, we are concerned with a different task, i.e., evaluat-

ing the suitability of LLMs to produce compressed summaries of Cochrane abstracts that remain suitable for healthcare professionals. To support critical decision making, such summaries must meet quality and acceptability criteria:

1. A high degree of **accuracy/factual correctness**. Hallucination remains one of the most significant issues in LLM-generated text (Zhang et al., 2025), and has the potential to cause significant harm in clinical decision making scenarios (Tang et al., 2023)
2. **Compliance with key aspects of PRISMA-A and Cochrane guidelines**, to ensure adequate coverage of relevant aspects of the systematic review.
3. **Professional acceptability**, meaning that compressed summaries must preserve clinically salient details required by experts.

Existing automated evaluation metrics (e.g., Lin (2004); Papineni et al. (2002); Lavie and Agarwal (2007); Zhang et al. (2019)) are insufficient to assess compliance with the above criteria, since they can only judge similarity between LLM-generated and gold-standard summaries at a general level. This motivates the development of more fine-grained and clinically-informed evaluation frameworks (Huang et al., 2025). While there is promising evidence in the general domain that LLMs could themselves replace humans as effective evaluators (Zheng et al., 2023; Dubois et al., 2024), this success does not yet extend to assessing medical evidence summaries automatically (Vasilev et al., 2026; Zhang et al., 2024; Joseph et al., 2024). Accordingly, human involvement in applying these frameworks remains irreplaceable.

We have thus developed a novel framework tailored to the evaluation of automatically-generated, expert-oriented summaries of Cochrane systematic review abstracts. The framework has subsequently been used to assess summaries generated by four different LLMs for 115 Cochrane abstracts in the mental health domain (i.e., a total 460 LLM-generated summaries). Evaluation was performed by two psychiatrists to ensure acceptability from a professional perspective. The assessment of each summary involves recording up to four pieces of information for each of nine different content dimensions, which are based on key requirements of the PRISMA-A and Cochrane handbook guidelines (Higgins et al., 2024).

2. Related Work

2.1. LLM-Generated Summaries in Biomedical and Clinical Domains

LLMs have demonstrated strong performance across a range of natural language processing tasks (Qin et al., 2024), including summarisation in the biomedical (Xie et al., 2023; Huang et al., 2025) and clinical (Bednarczyk et al., 2025) domains. Indeed, there is evidence that LLMs can sometimes outperform humans in accurately summarising medical evidence (Van Veen et al., 2024; Tang et al., 2023). Nevertheless, several concerns remain regarding the reliability of these summaries. Hallucination, i.e., the fabrication of details not present in the source text, remains a significant issue (Zhang et al., 2025). Biomedical and clinical summarisation pose additional challenges: many LLMs lack deep familiarity with specialised terminology and clinical concepts used in these domains (Huang et al., 2025), and even fine-tuned models can struggle to produce output that is sufficiently faithful and accurate (Zhang et al., 2024).

The risks become particularly acute in clinical decision-making contexts, where even seemingly small errors in the generated output could mislead readers, with potentially serious consequences (Kirmani et al., 2024; Jose et al., 2024). Tang et al. (2023) defined *harmfulness* as the potential for a generated summary to cause physical or psychological harm, or inappropriate changes in therapy, due to misinterpretation. Harmfulness in medical evidence summaries was primarily attributed to insufficient comprehensiveness (e.g., omission of Population, Intervention, Comparison, Outcome (PICO) elements (Richardson et al., 1995)) or to misinterpretation errors, including *certainty illusions*. These correspond to errors in the degree of certainty expressed towards particular findings, which could cause readers to under- or over-estimate their significance.

2.2. Automated Summary Evaluation

The high-stakes nature of LLM-generated medical summaries demands robust evaluation measures. Although general automated metrics (e.g., Lin (2004); Papineni et al. (2002); Lavie and Agarwal (2007); Zhang et al. (2019)) may suffice in non-critical scenarios, e.g., plain language summaries for non-experts (Arias-Russi et al., 2025a,b), they are inadequate for assessing compliance with clinically detailed criteria. Prior work demonstrates a failure of these metrics to reliably detect factual errors, hallucinations and harmfulness in medical summaries (Tang et al., 2023; Vasilev et al., 2025).

Such limitations motivate the development of domain-specific evaluation methods that explicitly

address terminological accuracy and clinical risk (Huang et al., 2025). While such frameworks are usually applied manually, there is growing interest in automation. The *LLM-as-judge* paradigm (Zheng et al., 2023) uses LLMs to evaluate model outputs across flexible criteria at lower cost and more efficiently than human application (Yu, 2025). Although promising in general domains (Zheng et al., 2023; Dubois et al., 2024), the effectiveness of this approach in detecting factual errors and hallucinations in medical summaries remains limited, particularly at the full-summary level (Vasilev et al., 2026; Zhang et al., 2024). However, LLM judges that evaluate finer-grained domain-specific aspects (e.g., PICO elements, side effects or medications) show stronger correlation with human judgments (Chu et al., 2025; Gero et al., 2024).

2.3. Human-Annotated Datasets

Although LLM-based evaluation of medical summaries remains imperfect, fine-tuning judge models on high-quality human-annotated data is a promising direction (Kim et al., 2023; Huang et al., 2025). Recent efforts to produce suitable datasets include Luo et al. (2024), Joseph et al. (2024) and Chu et al. (2025), which evaluate LLM-generated medical evidence summaries along multiple dimensions. These include a mixture of structural sections (e.g., *Objectives* and *Findings*) and specific content elements, such as PICO components, treatments and their effects, assessed through graded evaluation of aspects such as factuality, completeness and conciseness.

Joseph et al. (2024) further required evaluators to justify their scores through free-text rationales. These can support the development of *explainable* LLM judges, capable of articulating the basis for their evaluations. As base LLMs were found to generate weak explanations, human-authored rationales provide valuable supervisory data to improve evaluative reasoning. These higher-quality critiques may also support iterative refinement of generated summaries (Saunders et al., 2022).

3. Methods

3.1. Evaluation Framework

Existing evaluation frameworks for medical summarisation provide useful foundations, but do not fully address the requirements of assessing professionally-oriented systematic review summaries. To bridge this gap, we have developed a structured evaluation framework aligned with aspects of the PRISMA-A and Cochrane guidelines. The design of the framework targets the assessment not only of factual accuracy, but also of pro-

fessional acceptability in compressed summaries intended for healthcare decision makers.

Each summary is evaluated across nine different dimensions reflecting key structural and evidential components of systematic review abstracts. For each dimension, evaluators assign a single ordinal score capturing factual adequacy and professional suitability. Certain dimensions with identified deficiencies additionally require the specification of the type of corrective action that would be needed to achieve content acceptability (i.e., *add*, *remove* or *modify*). This information enables a systematic analysis of common error types.

To ensure interpretability and traceability of errors (Kambhamettu et al., 2025; Xie et al., 2024), non-maximal scores require erroneous content in generated summaries to be linked to corresponding evidence in the source abstract. Specifically, problematic sentences in the summary are identified at the dimension level, while accompanying dimension-specific rationales explain the issues.

The nine evaluation dimensions are outlined below, with scoring details provided in Table 1. Appendix A provides an example of how the evaluation scheme is applied to a specific summary.

Dimensions 1–4: PICO elements. PRISMA-A guidelines require systematic review abstracts to specify study objectives in terms of PICO elements (see Table 2). Since Tang et al. (2023) identified missing PICO details as a source of harmfulness in LLM-generated medical summaries, we evaluate *Population*, *Intervention*, *Comparator* and *Outcome* elements as separate dimensions.

Dimensions 5–7: Methodology, Main Findings and Conclusions. These reflect PRISMA-A requirements that abstracts should report study methods, key findings and a concluding interpretation, including possible limitations.

Dimension 8: Quality of Evidence. The Cochrane handbook specifies that evidence syntheses should report a GRADE assessment of evidence quality/certainty (Balslem et al., 2011), i.e., *high*, *moderate*, *low* or *very low*. This dimension assesses whether the generated summary accurately reflects the quality of evidence reported in the original abstract.

Dimension 9: Currency. PRISMA-A requires abstracts to report when information sources were last searched. This dimension evaluates whether the generated summary correctly represents the search date(s) specified in the abstract.

3.2. Data Source

Our evaluation corpus, which we name *Psyc-SumEval*, was constructed using the abstracts of 115 randomly selected, publicly available Cochrane systematic reviews in the mental health domain. Reviews were sampled from the

Dimension	Scoring scheme and criteria
PICO elements (Population, Intervention, Comparator, Outcome)	0 = missing; 1 = partially correct or inaccurate; 2 = accurate and complete. Rationale required for 0/1.
Quality of Evidence	0 = inconsistent with abstract or only partially reported; 1 = consistent with abstract (both present or both absent). Rationale required for 0.
Main Findings	0 = missing; 1 = partial/inaccurate; 2 = accurate. Experts also specify corrective action: add, remove, modify or none. Rationale required for 0/1
Methodology	Same 0–2 scale and corrective action scheme as Main Findings.
Conclusions	Same 0–2 scale and corrective action scheme as Main Findings.
Currency	0 = date inaccurate or omitted; 1 = date accurate; “No date reported” = absent in both abstract and summary.

Table 1: Rubric used to evaluate LLM-generated summaries

PICO Element	Description	Examples
Population	Characteristics of trial participants (e.g, demographics, medical condition(s))	<i>Adult (19–44 years), Schizophrenia</i>
Intervention	Active treatment assessed for efficacy	<i>Family Therapy, Relapse Prevention</i>
Comparator	Alternative treatment against which the intervention is compared	<i>Usual care</i>
Outcome	Measures used to determine trial results	<i>Hospital admission, Quality of life</i>

Table 2: Descriptions and examples of PICO elements

Cochrane Library¹, to ensure diversity of interventions and populations. No patient-level or individual trial data were included.

3.3. Summary Generation

For each of the 115 abstracts, a prompt designed to encourage the generation of a compressed summary compliant with key PRISMA-A and Cochrane guidelines (see Figure 1) was issued to four different LLMs, i.e., *GPT-4o* (OpenAI, 2024a), *GPT-4o-mini* (OpenAI, 2024b), *Claude Opus* (Anthropic, 2024) and *LLaMA-3.1-70B-Instruct* (Meta, 2024). No content beyond the abstract itself was provided to the models. This yielded 460 summaries, each of which evaluated by one of the two psychiatrists using the framework described above.

3.4. Evaluation of Professional Judgement Stability

The primary evaluation phase was followed by a cross-review phase. The objective was not to establish inter-annotator reliability in the traditional sense, but rather to examine the stability of professional judgment under a shared, structured rubric. In high-stakes biomedical summarisation, variation between experts may reflect legitimate interpretive discretion rather than labelling error. The

Summarize the following research abstract using the PICO framework. Identify the Population (P) being studied, the Intervention (I), the Comparison (C), and Outcomes (O) measured. Highlight the study's methodology, main findings, and conclusions. Emphasize any key clinical outcomes, the quality of the evidence, and how up-to-date the evidence is. Ensure the summary captures the overall significance of the research in the context of patient care and treatment efficacy, as well as any potential limitations or recommendations for future research.

Figure 1: Prompt issued to LLMs

cross-review phase was therefore designed to permit a cross-dimensional analysis of convergence and divergence in professionally grounded perspectives on acceptable abstraction of evidence, rather than to produce a statistically optimised annotation dataset. As mentioned above, every non-maximal score required an explanatory rationale, linking evaluative outcomes to explicitly documented reasoning. Accordingly, disagreement was analysed at the level of justification, rather than being restricted to numerical divergence.

Although Cohen's κ is commonly used to quantify inter-annotator agreement beyond chance, it was not adopted here. κ assumes independent annotation over fully overlapping datasets with

¹<https://www.cochranelibrary.com/>

fixed categorical labels and treats disagreement as noise relative to a latent ground truth. However, we aimed to examine professional judgment stability rather than categorical reliability. Moreover, several dimensions involve graded or interpretive assessments (e.g., adequacy of methodological detail or appropriateness of conclusions), where variation may reflect professional discretion. Under these conditions, a κ coefficient could conflate meaningful interpretive differences with unreliability. We therefore report dimension-level disagreement patterns alongside analysis of qualitative rationales to characterise expert judgment stability.

The cross-review phase was undertaken using only a subset of the complete corpus (i.e., 30 randomly selected abstracts), in order to balance analytical depth with realistic constraints on expert time, prioritising dimension-level interpretability over large-scale redundancy. Each expert reviewed the four generated summaries for 15 abstracts originally annotated by the other expert (60 summaries total) and indicated agreement or disagreement for each dimension.

4. Results

4.1. Cross-Review Phase

Table 3 reports the number of disagreements between the psychiatrists identified in the cross-review phase. Patterns of disagreement provide insight into which aspects of summarisation are primarily governed by explicit guideline adherence and which aspects are more open to interpretive professional discretion. For most models and dimensions, disagreements occurred in a relatively small proportion of cases (generally $\leq 15\%$ of the 30 summaries reviewed). This provides evidence of substantial convergence between experts regarding professional acceptability across dimensions. A notable exception is the *Conclusions* dimension for *LLaMA-3.1-70B*, whose closing content was often repetitive and poorly structured, making consistent evaluation difficult. We provide a dimension-level analysis below.

PICO Elements — The *Objectives* sections of original abstracts often specify PICO elements at a broad, review design level, while the *Main Results* may provide more detailed PICO characteristics that are specific to the studies finally included in the review (e.g., age ranges of study populations or names of specific interventional drugs). The two experts sometimes disagreed over the necessity of including such study-specific details in compressed summaries. In the original abstracts, these additional details are most frequently provided for the *Population* and *Intervention* elements, but were most commonly omitted in the summaries

of *LLaMA-3.1-70B-Instruct* and *GPT-4o-mini*, thus explaining the concentration of disagreements in these two dimensions and models.

Methodology — Experts showed high agreement regarding core methodological information, particularly the inclusion of total numbers of studies and participants. Disagreements questioned the need to include more granular information, such as specific registers searched (e.g., *Cochrane Schizophrenia study-based register of trials*), countries covered, or the level of detail required to describe data collection and analysis, and risk-of-bias assessment.

Main Findings — Most disagreements concerned models' handling of low-certainty evidence. In particular, there was some difference of opinion on whether results specified as having *very low* certainty in the original abstract could be safely omitted in the compressed summary. Variation also emerged in terms of tolerance towards subtle shifts in certainty statements. For example, the original abstract statement *None of the trials reported clinically important change in quality of life* was rendered by one model as *Modafinil has no effect on quality of life*, which only one expert accepted. In contrast, more hedged formulations (e.g., *may have little or no effect*) were consistently judged to be acceptable.

Conclusions — Some model-generated summaries separated *Conclusions*, *Limitations* and *Recommendations for Future Research*. One expert assessed *only* the explicitly labelled *Conclusions* sections, whereas the other expert also considered content in *Limitations* and *Recommendations*, given that such information may occur within original abstract conclusions. Disagreements typically arose when the latter expert found potentially misleading or unsupported statements in these additional sections. If the original abstract contained little or no discussion of limitations or future research directions, then all models were susceptible to introducing new speculative content without explicit attribution. Notably, however, no disagreements concerned the content of explicitly labelled *Conclusions* sections of the generated summaries, suggesting that the experts had shared expectations regarding the core function of these sections.

Quality of Evidence & Currency — Disagreements for *Quality of Evidence* were confined to a single abstract involving an atypical case, in which no eligible studies were identified. The sole disagreement for *Currency* appears attributable to evaluator oversight, as the search date was found to be explicitly stated in the generated summary.

4.2. Per-Model Performance

Using the psychiatrist-assigned assessment scores, we compared model performance across

Category	GPT-4o	Claude Opus	LLaMA-3.1-70B	GPT-4o-mini
Population	2	3	4	4
Intervention	1	2	5	2
Comparator	0	0	1	1
Outcome	0	0	0	0
Methodology	2	2	2	3
Main Findings	4	3	5	4
Conclusions	5	5	13	6
Quality of Evidence	1	1	1	1
Currency	0	1	0	0

Table 3: Number of disagreements in the cross-checked subset ($n = 30$ reviews per model).

Dimension	GPT-4o	Claude Opus	LLaMA-3.1-70B	GPT-4o-mini
Population (0–2)	1.91	1.77	1.71	1.70
Intervention (0–2)	1.87	1.83	1.79	1.80
Comparator (0–2)	1.90	1.92	1.81	1.90
Outcome (0–2)	1.84	1.86	1.77	1.76
Methodology (0–2)	1.65	1.43	1.28	1.42
Main Findings (0–2)	1.74	1.77	1.37	1.70
Conclusions (0–2)	1.72	1.72	1.30	1.57
Quality of Evidence (1)	0.93	0.96	0.98	0.97
Currency (1)	0.41	0.39	0.97	0.77

Table 4: Per-model performance statistics. Means (0–2 scale) are provided for ordinal dimensions and proportion of “1” labels is shown for Quality of Evidence and Currency.

dimensions. For dimensions evaluated on a 0–2 scale, mean scores were calculated to provide an indication of overall model adequacy. For the dimensions with binary values (i.e., *Quality of Evidence* and *Currency*), we report the proportion of summaries assigned the value “1”. Although *Currency* allowed three labels (“0”, “1” and “No date reported”), only the first two labels were ever assigned. Results are presented in Table 4.

Table 5 reports the distribution of *corrective actions* needed to render the content of the *Methodology*, *Main Findings* and *Conclusions* sections acceptable. These actions are almost exclusively *Add* or *Remove*, showing that most content errors in generated summaries correspond to omissions or hallucinated/unsupported content, as discussed further below. The tiny proportion of *Modify* actions suggests that misinterpretations/distortions of original abstract content are rare.

4.2.1. PICO Dimensions

Performance on extracting PICO elements is generally high. Although expert rationales occasionally note completely missing items in generated summaries, the most common issue concerns mismatches between the required and provided levels of detail. Although a small number of disagreements regarding the required level of detail in PICO element summaries were identified in Section 4.1, the two experts generally agreed

that such summaries should include both *broader scope characteristics* defined in the original review objectives and more *specific characteristics* of the studies actually included in the review. Errors in PICO elements often correspond to cases in which one or other of these types of characteristics is missing from the generated summary.

When only a single study is included in a review, for example, models often extract specific PICO characteristics from that study alone, and omit broader scope review objectives. However, for reviews including multiple studies, it is more common for summaries to include these broader characteristics, but to omit additional study-specific details. This pattern of behaviour is most prevalent for *Population*, although the notably higher performance of *GPT-4o* for this dimension reflects its greater tendency to generate *Population* summaries that include not only general information from the *Objectives* section of the original abstract (e.g., conditions suffered), but also more specific details from the *Selection Criteria* and *Main Results* sections, such as age, ethnicity and condition severity.

More stable cross-model performance is observable for *Intervention*, with all models demonstrating an ability to extract relevant information from the *Objectives*, *Selection Criteria* and *Main Results* sections of the original abstract. This allows generated summaries to successfully supplement, e.g., classes of interventional drugs introduced in the

Dimension	Corrective Action	GPT-4o	Claude Opus	LLaMA-3.1-70B	GPT-4o-mini
Methodology	None	65.2%	42.6%	27.8%	41.7%
	Add	33.0%	55.7%	71.3%	58.3%
	Remove	0.0%	0.9%	0.0%	0.0%
	Modify	1.7%	0.9%	0.9%	0.0%
Main Findings	None	74.8%	76.5%	37.4%	70.4%
	Add	25.2%	21.7%	60.9%	29.6%
	Remove	0.0%	0.0%	0.0%	0.0%
	Modify	0.0%	1.7%	1.7%	0.0%
Conclusions	None	71.3%	73.0%	30.4%	57.4%
	Add	0.9%	0.9%	0.9%	0.9%
	Remove	27.8%	26.1%	68.7%	41.7%
	Modify	0.0%	0.0%	0.0%	0.0%

Table 5: Distribution of corrective actions for *Methodology*, *Main Findings* and *Conclusions*

objectives with the names of more specific drugs evaluated in the included studies. The slightly higher performance of *GPT-4o* for this dimension appears to be due to its stronger ability to extract specific treatments when their mentions are scattered throughout the *Main Results* section of the original abstract.

Performance for *Comparator* is high for all models. However, when multiple comparators are mentioned in the original abstract, *LLaMA-3.1-70B* appears to be slightly more prone than other models to omitting one or more of these comparators from the generated summary.

In certain original abstracts, some or all of the *Outcomes* are mentioned only in the *Main Results* sections. Because these sections are often information-dense, comprehensive outcome extraction can be more challenging than other elements. The larger *GPT-4o* and *Claude Opus* handle this complexity more consistently, which likely explains their slightly stronger performance for the *Outcome* dimension.

4.2.2. Methodology Dimension

This dimension exhibits the lowest performance scores across all models. Achieving acceptability is reliant on integrating selected information scattered across multiple sections of original abstracts (typically *Search Methods*, *Selection Criteria*, *Data Collection and Analysis* and *Main Results*). This makes appropriate content selection challenging.

Expert rationales indicate strong agreement on a core set of expected elements, i.e., whether the review updates a previous one; key resources searched; number of included studies; total participant numbers; and brief descriptions of data analysis, risk-of-bias assessment and evidence grading. Table 5 shows that *Methodology* errors almost exclusively correspond to omission of certain information that is considered necessary from a professional perspective. Usually, however, the deficiency stems from a single missing core item,

rather than more widespread inaccuracies.

Inspection of generated summaries reveals that all models are capable of extracting information relevant to multiple core elements, but they do so inconsistently. This may reflect a tendency to generate relatively fixed-length *Methodology* sections, in which the inclusion of less critical details displaces more essential ones. Consequently, fundamental information, such as the number of included studies, may be omitted. Compared to the stronger *GPT-4o*, this pattern is more evident for *Claude Opus*, *LLaMA-3.1-70B-Instruct* and *GPT-4o-mini*, whose methodological summaries frequently restate PICO information narratively and/or redundantly include the search date, even when it appears elsewhere in the summary.

4.2.3. Main Findings Dimension

Generating acceptable summaries of *Main Findings* may appear simpler than for *Methodology*, partly because this task primarily involves compressing a single section of the original abstract, i.e., *Main Results*, and also since the performance for all models is higher than for *Methodology*. Nevertheless, the task remains complex: original *Main Results* sections are often dense, given their integration of results, interpretations and certainty levels across multiple PICO permutations. Effective summarisation thus requires both accurate extraction and disciplined compression.

This complexity helps to explain why the smaller *LLaMA-3.1-70B-Instruct* struggles with this dimension. Table 5 shows that, similarly to *Methodology*, most identified issues concern omissions of necessary information. Contrary to concerns raised in the cross-review analysis, overt *certainty illusions* are rare. More frequent is the failure to explicitly associate reported results with their originally stated level of evidence quality/certainty.

By far the most common problem, however, is the omission of one or more outcomes mentioned in the original abstract. While one expert was

sometimes slightly more lenient, both experts were largely unanimous that *all* outcomes identified in the *Outcomes* dimension should be mentioned in the *Main Findings* of the compressed summary, regardless of evidence quality level and whether sufficient data was available to evaluate the outcome.

Table 5 indicates that high-quality information extraction and compression appears to be feasible for all models, with the exception of *LLaMA-3.1-70B-Instruct*. Different models use varying techniques to achieve the goal of generating *Main Findings* summaries that are shorter than the corresponding section in the original abstract. One method that appears to be used to varying degrees by all models is the implicit application of thresholding criteria to filter out some of the originally reported outcomes. Examples include secondary outcomes; outcomes qualified by phrases like *very low certainty*, *no clear evidence* or *interpreted with caution*; outcomes with no usable data; and cases in which the intervention(s) evaluated made little or no difference to specified outcome(s). However, given the experts' opinion that all outcomes from the original abstract should also be mentioned in the generated summaries, such filtered results lead to lower acceptability scores.

As shown in Table 5, *Claude Opus* and *GPT-4o* generate the highest proportions of *Main Findings* summaries that are considered to be acceptable without modification (76.5% and 74.8%, respectively). This likely reflects their stronger abstractive capabilities, enabling compact integration of multiple results with minimal loss of information. The slightly lower performance of the smaller *GPT-4o-mini* (70.4% of *Main Results* sections deemed acceptable) may reflect its tendency to adhere more closely to the more verbose source phrasing of results, meaning that a greater degree of outcome filtering may sometimes be required to produce *Main Findings* summaries of a suitable length. The far lower proportion of fully acceptable *Main Findings* generated by *LLaMA-3.1-70B-Instruct* appears to be due to its frequent generation of shorter (and hence, more extremely filtered) findings sections than other models, possibly due to the disproportionate length allocated to its concluding content.

4.2.4. Conclusions Dimension

Performance on the *Conclusions* dimension is lower than for *Main Findings* across all models. This may seem surprising, since *Conclusions* sections from the original abstracts should be more straightforward to compress than the more detailed *Main Results* sections; conclusions are uniformly shorter, and aim to clearly summarise the most significant outcomes of the review.

Indeed, the expert who evaluated only explicitly labelled *Conclusions* sections judged the vast ma-

jority of these generated sections to be acceptable without modification. The main exception was for *LLaMA-3.1-70B-Instruct*, whose conclusions were frequently marked down for repetition of information.

The lower scores for this dimension mainly stem from problems in the generated content of sections labelled *Limitations*, *Future Research* or similar, which only the second expert considered. The issues were mostly the same as those identified in the cross-review analysis, i.e., fabricated content within these additional sections, which was particularly likely to occur if the original abstract provided little or no information about limitations or future work. In such cases, models often generated statements that appeared plausible, but which were unsupported by the original abstract content. This pattern was especially pronounced in the outputs of *LLaMA-3.1-70B-Instruct* (and to a lesser extent, *GPT-4o-mini*), which sometimes framed invented content as directly attributable to the study (e.g., *The study highlights ...*). Such observations help to explain why the *Corrective Action* specified for the vast majority of erroneous *Conclusions* sections is to *remove* information (see Table 5).

However, given that the second expert identified few issues within explicitly labelled *Conclusions* sections, all models appear generally proficient at generating accurate summaries of primary study outcomes. Similarly to *Main Findings*, preserving nuanced expressions of certainty within the *Conclusions* sections remains important to avoid misinterpretation, and expert rationales indicate that fidelity to certainty levels is generally maintained.

4.2.5. Quality of Evidence Dimension

Results show that the vast majority of generated summaries preserve sufficient information from original abstracts regarding evidence quality. Adherence to Cochrane Handbook guidelines regarding GRADE evidence certainty levels usually results in uniform specification of these levels across original abstracts, thus making this information straightforward to extract. Examination of generated summaries reveals variations in exactly how and where information about quality of evidence is specified. This may, for example, be at the level of individual results in the *Main Findings* section; within a separate *Quality of Evidence* section that specifies the range of certainty levels across all results; or as a statement in the *Limitations* section that low or very low evidence quality limits the validity of the results. All of these reporting methods appear to be professionally acceptable, as long as the complete range of evidence levels observed across all results is specified. Only a small number of summaries fail to meet this criterion.

4.2.6. Currency Dimension

The results for this dimension are unexpected. In contrast to other dimensions, the smaller *LLaMA-3.1-70B-Instruct* and *GPT-4o-mini* models considerably outperform the larger *GPT-4o* and *Claude Opus*. Particularly noteworthy is *LLaMA-3.1-70B-Instruct* which, despite under-performing other models in all other dimensions, is able to correctly extract information about search currency from 97% of abstracts, compared to only around 40% of abstracts for *GPT-4o* and *Claude Opus*. Extraction of this information from original abstracts should theoretically be very straightforward, since search dates are usually clearly stated in the *Search Methods* section. Therefore, the failure of larger models to extract this information consistently is surprising. A feasible explanation could be the wording of the prompt issued to the LLMs, which explicitly elicits information about how up to date the search *evidence* is. It is therefore possible that other models are carrying out more complex processing to ensure that a search date is *only* included in the generated summary if it is clear that all available evidence up until the specified search date was considered. As an example, some original abstracts state that a search was carried out *without date restrictions*. However, further investigation is needed to determine whether different models' decisions to exclude search dates from summaries is based on prompt wording and/or contextual cues, or is simply due to space constraints.

5. Discussion

We have shown that our proposed multi-dimensional framework can be applied by clinical experts to evaluate LLM-generated summaries from a professional perspective. The cross-review results indicate substantial convergence in expert judgments, suggesting that shared professional norms guide evaluation across dimensions. The evaluations enabled systematic comparison of model performance. No single model outperformed the others across all dimensions; instead, performance was dimension-specific. Analysis of corrective actions revealed consistent patterns: deficiencies in *Methodology* and *Main Findings* largely involved omissions of essential details, whereas problematic *Conclusions* more often contained unsupported or fabricated content. The frequency of such issues varied by model.

Expert-authored rationales provided further insight into implicit professional criteria, including core methodological elements expected in summaries and the requirement that all pre-specified outcomes be acknowledged in *Main Findings*, even when evidence is inconclusive. These expectations are not always explicit in reporting guide-

lines but are central to professional acceptability. Capturing such implicit standards highlights the value of structured expert annotation for improving both evaluation and generation of automatically generated systematic review summaries.

Our findings also reveal that different dimensions pose distinct challenges. For *PICO* elements, the key issue is balancing design-level scope with study-specific detail. *Methodology* requires selective integration of key information that may be scattered across multiple abstract sections. *Main Findings* demand effective abstraction to preserve *all* outcomes within limited space, while *Conclusions* require strict fidelity to the source text to avoid unsupported extrapolation or hallucination. To meet professional expectations for *Quality of Evidence*, the complete range of evidence levels observed across reported results must be present in the generated summary. Finally, the unexpectedly poor results of the *Currency* dimension for most models demand further investigation to ensure that models' understanding is aligned with expert requirements for this dimension; the latter appear to be restricted to the need for a simple specification of the date(s) when resources relevant to the review were searched.

6. Conclusion

This paper has introduced a novel framework for evaluating the professional acceptability of LLM-generated summaries of systematic review abstracts in high-stakes clinical settings. The framework has been applied by psychiatrists to assess summaries generated by four different LLM models for 115 Cochrane abstracts to create the *PsycSumEval* corpus. Beyond benchmarking model performance, *PsycSumEval* makes expert evaluative reasoning explicit through structured scoring and documented rationales. In scenarios where summaries may influence clinical understanding and decision making, interpretability and transparency of professional judgement is essential for responsible evaluation.

7. Future Work

Future work will explore the effects of alternative prompting strategies on model performance. We also plan to extend our work beyond the mental health domain to systematic reviews in other clinical areas, improving coverage and generalisability. Involving additional clinical experts will allow us to examine the stability of professional acceptability criteria across domains. Furthermore, we will use *PsycSumEval* to fine-tune and evaluate LLM judges tailored to professionally oriented assessment of systematic review summaries.

8. Limitations

Several limitations should be noted. The evaluation involved only two psychiatrists, potentially restricting diversity of perspectives. The cross-review was conducted on a subset of 30 reviews (120 summaries), while the remaining summaries were evaluated by a single expert. Broader double evaluation and the involvement of additional experts would strengthen generalisability. The cross-review protocol was not fully blinded, as experts had access to prior labels and rationales. While this supported interpretability, it may have introduced mild anchoring effects. Future work should adopt and compare fully independent evaluations. The study assessed summaries of Cochrane abstracts, rather than full systematic reviews. Although abstracts provide a structured and controlled setting, they represent only part of the underlying evidence. Extending evaluation to full-text reviews may yield different performance patterns.

9. Ethics Statement

The PsycSumEval corpus is derived from publicly available Cochrane abstracts and contains no patient-level or personally identifiable data. However, it involves clinically sensitive content: summaries of medical evidence may influence decisions if misused. The resource is therefore released solely for methodological research and not for clinical application. Risks include reinforcement of model errors in high-stakes settings, particularly in interpretive dimensions such as conclusions. Continued human oversight remains essential. Although abstracts include review dates, medical evidence evolves and findings may become outdated. All evaluation protocols and guidelines are released transparently to support responsible and reproducible research.

10. Data Availability

The PsycSumEval corpus and evaluation guidelines are available at <https://github.com/paul-rich-t/PsycSumEval>.

11. Acknowledgements

This work was partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union under the NextGenerationEU Program.

12. Bibliographical References

Gerd Antes, David Tovey, and Nancy Owens. 2022. Cochrane: Trusted evidence. informed decisions. better health. *Systematic Reviews in Health Research: Meta-Analysis in Context*, pages 413–423.

Felipe Arias-Russi, Carolina Salazar-Lara, and Rubén Manrique. 2025a. [Bridging the gap in health literacy: Harnessing the power of large language models to generate plain language summaries from biomedical texts](#). In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 269–284, Albuquerque, New Mexico. Association for Computational Linguistics.

Felipe Arias-Russi, Carolina Salazar Lara, and Ruben Manrique. 2025b. [A multi-agent framework with diagnostic feedback for iterative plain language summary generation from cochrane medical abstracts](#). In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 87–104, Suzhou, China. Association for Computational Linguistics.

Howard Balshem, Mark Helfand, Holger J. Schünemann, Andrew D. Oxman, Regina Kunz, Jan Brozek, Gunn E. Vist, Yngve Falck-Ytter, Joerg Meerpohl, Susan Norris, and Gordon H. Guyatt. 2011. [GRADE guidelines: 3. Rating the quality of evidence](#). *Journal of Clinical Epidemiology*, 64(4):401–406.

Lydie Bednarczyk, Daniel Reichenpfader, Christophe Gaudet-Blavignac, Amon Kenna Ette, Jamil Zaghir, Yuanyuan Zheng, Adel Bensahla, Mina Bjelogrić, and Christian Lovis. 2025. Scientific evidence for clinical text summarization using large language models: scoping review. *Journal of Medical Internet Research*, 27:e68998.

Elaine M Beller, Paul P Glasziou, Douglas G Altman, Sally Hopewell, Hilda Bastian, Iain Chalmers, Peter C Gøtzsche, Toby Lasserson, David Tovey, and PRISMA for Abstracts Group. 2013. PRISMA for abstracts: reporting systematic reviews in journal and conference abstracts. *PLoS medicine*, 10(4):e1001419.

Elaine M Beller, Paul P Glasziou, Sally Hopewell, and Douglas G Altman. 2011. Reporting of effect direction and size in abstracts of systematic reviews. *Jama*, 306(18):1981–1982.

- BE Belsher, E Beech, D Evatt, DJ Smolenski, MT Shea, JL Otto, CS Rosen, and PP Schnurr. 2019. [Present-centered therapy \(pct\) for post-traumatic stress disorder \(ptsd\) in adults](#). *Cochrane Database of Systematic Reviews*, (11).
- Iain Chalmers and Douglas G Altman. 1995. *Systematic reviews*. BMJ Publishing London.
- Bohao Chu, Meijie Li, Sameh Frihat, Chengyu Gu, Georg Lodde, Elisabeth Livingstone, and Norbert Fuhr. 2025. [TracSum: A new benchmark for aspect-based summarization with sentence-level traceability in medical domain](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 844–864, Suzhou, China. Association for Computational Linguistics.
- Declan Devane, Johanna Pope, Paula Byrne, Evan Forde, Isabel O'Byrne, Steven Woloshin, Eileen Culloty, Darren Dahly, Ingeborg Hess Elgersma, Heather Munthe-Kaas, Conor Judge, Martin O'Donnell, Finn Krewer, Sandra Galvin, Nikita N. Burke, Theresa Tierney, KM Saif-Ur-Rahman, Tom Conway, and James Thomas. 2026. [Comparison of AI-assisted and human-generated plain language summaries for Cochrane reviews: a randomised non-inferiority trial \(HIET-1\) \[Registered Report - stage II\]](#). *Journal of Clinical Epidemiology*, 191:112102.
- Kay Dickersin, Roberta Scherer, and Carol Lefebvre. 1994. Identifying relevant studies for systematic reviews. *BMJ*, 309(6964):1286–1291.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.
- Zelalem Gero, Chandan Singh, Yiqing Xie, Sheng Zhang, Praveen Subramanian, Paul Vozila, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2024. Attribute structuring improves LLM-based evaluation of clinical text summaries. *arXiv preprint arXiv:2403.01002*.
- Jasmin Helbach, Falk Hoffmann, Dawid Pieper, and Katharina Allers. 2023a. Reporting according to the preferred reporting items for systematic reviews and meta-analyses for abstracts (PRISMA-A) depends on abstract length. *Journal of Clinical Epidemiology*, 154:167–177.
- Jasmin Helbach, Kathrin Wandscher, Dawid Pieper, and Falk Hoffmann. 2023b. Cochrane review abstracts are getting longer, but this has no large impact on the reporting quality. *Journal of Evidence-Based Medicine*, 16(3):294–302.
- JPT Higgins, J Thomas, J Chandler, M Cumpston, T Li, MJ Page, and VA Welch. 2024. [Cochrane Handbook for Systematic Reviews of Interventions version 6.5 \(updated August 2024\)](#). Cochrane.
- Zhenyu Huang, Xianlai Chen, Yunbo Wang, Jincan Huang, and Xing Zhao. 2025. [A survey on biomedical automatic text summarization with large language models](#). *Information Processing & Management*, 62(5):104216.
- Sagar Jose, Khanh T.P Nguyen, Kamal Medjaher, Ryad Zemouri, Mélanie Lévesque, and Antoine Tahan. 2024. [Advancing multimodal diagnostics: Integrating industrial textual data and domain knowledge with large language models](#). *Expert Systems with Applications*, 255:124603.
- Sebastian Joseph, Lily Chen, Jan Trienes, Hannah Göke, Monika Coers, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2024. [FactPICO: Factuality evaluation for plain language summarization of medical evidence](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8437–8464, Bangkok, Thailand. Association for Computational Linguistics.
- Hita Kambhamettu, Jamie Flores, and Andrew Head. 2025. Traceable texts and their effects: A study of summary-source links in ai-generated summaries. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.
- Mahira Kirmani, Gagandeep Kour, Mudasir Mohd, Nasrullah Sheikh, Dawood Ashraf Khan, Zahid Maqbool, Mohsin Altaf Wani, and Abid Hussain Wani. 2024. Biomedical semantic text summarizer. *BMC Bioinformatics*, 25(1):152.
- Alon Lavie and Abhaya Agarwal. 2007. [ME-TOR: An automatic metric for MT evaluation with high levels of correlation with human judgments](#). In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.
- Alessandro Liberati, Douglas G Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C Gøtzsche, John PA Ioannidis, Mike Clarke, Philip J Devereaux, Jos Kleijnen, and David Moher. 2009.

- The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*, 339.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2024. [Factual consistency evaluation of summarization in the era of large language models](#). *Expert Systems with Applications*, 254:124456.
- Katarina Maticic, Marina Krnic Martinic, and Livia Puljak. 2019. Assessment of reporting quality of abstracts of systematic reviews with meta-analysis using PRISMA-A and discordance in assessments between raters without prior experience. *BMC Medical Research Methodology*, 19(1):32.
- Sally Morton, Alfred Berg, Laura Levit, and Jill Eden. 2011. *Finding what works in health care: standards for systematic reviews*. National Academies Press.
- Ian G Needleman. 2002. A guide to systematic reviews. *Journal of clinical periodontology*, 29:6–9.
- Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021a. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.
- Matthew J Page, David Moher, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021b. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372.
- 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.
- Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet NLP: A survey. *arXiv preprint arXiv:2405.12819*.
- W Scott Richardson, Mark C Wilson, Jim Nishikawa, and Robert S Hayward. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123(3):A12–13.
- William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.
- Liyan Tang, Zhaoyi Sun, Betina Ilnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *NPJ Digital Medicine*, 6(1):158.
- Amy Y Tsou and Jonathan R Treadwell. 2016. Quality and clarity in systematic review abstracts: an empirical study. *Research synthesis methods*, 7(4):447–458.
- Dave Van Veen, Cara Van Uden, Louis Blanke-meier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142.
- Yuriy Vasilev, Irina Raznitsyna, Anastasia Pamova, Tikhon Burtsev, Tatiana Bobrovskaya, Pavel Kosov, Anton Vladzimirskyy, Olga Omelyanskaya, and Kirill Arzamasov. 2025. Evaluating medical text summaries using automatic evaluation metrics and LLM-as-a-judge approach: A pilot study. *Diagnostics*, 16(1):3.
- Yuriy Vasilev, Irina Raznitsyna, Anastasia Pamova, Tikhon Burtsev, Tatiana Bobrovskaya, Pavel Kosov, Anton Vladzimirskyy, Olga Omelyanskaya, and Kirill Arzamasov. 2026. [Evaluating medical text summaries using automatic evaluation metrics and llm-as-a-judge approach: A pilot study](#). *Diagnostics*, 16(1).
- Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. 2023. A survey for biomedical text summarization: From pre-trained to large language models. *arXiv preprint arXiv:2304.08763*.
- 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 evaluation for medical text generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 649–679, Bangkok, Thailand. Association for Computational Linguistics.

Fangyi Yu. 2025. When AIs judge AIs: The rise of agent-as-a-judge evaluation for LLMs. *arXiv preprint arXiv:2508.02994*.

Gongbo Zhang, Qiao Jin, Yiliang Zhou, Song Wang, Betina Ilnay, Yiming Luo, Elizabeth Park, Jordan G Nestor, Matthew E Spotnitz, Ali Soroush, et al. 2024. Closing the gap between open source and commercial large language models for medical evidence summarization. *NPJ Digital Medicine*, 7(1):239.

Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2025. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 46595–46623, Red Hook, NY, USA. Curran Associates Inc.

shows the summary of this abstract, generated by *GPT-4o*, whose length is around half that of the original abstract (i.e., 422 words).

Table 6 shows the expert-provided scores and rationales (for dimensions whose content is not considered to be fully professionally acceptable) for the *Population*, *Intervention*, *Comparator*, *Outcome*, *Quality of Evidence* and *Currency* dimensions. Meanwhile, Table 7 shows the expert-provided scores, corrective actions and rationales (for dimensions whose content is not considered to be fully professionally acceptable) for the *Methodology*, *Main Findings* and *Conclusions* dimensions. For all dimensions in each table, a short discussion explaining the assigned score is provided, by making comparisons between the content of the original abstract and the generated summary. Note that these discussions were not provided by the experts as part of their evaluation effort, but rather have been added to aid understanding of the evaluation process.

The evaluation results shows that *GPT-4o* was largely successful in generating a significantly compressed summary of the original abstract, which largely conforms to the required standards of professional acceptability. Although the final sentence of the generated summary was found to constitute hallucinated content that is not supported by the original abstract, only a small number of required details were found to be missing from the remainder of the summary.

13. Language Resource References

Anthropic. 2024. *Claude Opus*.

Meta. 2024. *LLaMA 3.1 70B Instruct*.

OpenAI. 2024a. *GPT-4o*.

OpenAI. 2024b. *GPT-4o-mini*.

A. Worked Evaluation Example

This appendix provides a detailed, worked example of how a specific summary, generated by *GPT-4o*, was evaluated by one of the psychiatrist experts against the original Cochrane abstract (Belsher et al., 2019) (ID: *CD012898.PUB2*) according to the nine dimensions of our evaluation scheme.

Figure 2 provides the text of the original abstract, which is 865 words long. Meanwhile, Figure 3

Dimension	Score	Rationale	Discussion
Population	2	-	Relevant population information (i.e., <i>adults with PTSD</i>) is specified in <i>Objectives</i> section of the original abstract, and was correctly extracted. In the <i>Population</i> element of the generated summary, information from the <i>Background</i> section of the original abstract has also been used to expand <i>PTSD</i> to the full form <i>post-traumatic stress disorder</i> .
Intervention	2	-	Relevant information about the single intervention of interest (i.e., <i>PCT</i>) is correctly extracted from the <i>Objectives</i> section of original abstract, and is supplemented in the generated abstract with both the full form, i.e., <i>Present-centered therapy</i> and additional descriptive information, both of which are provided in the <i>Background</i> section of the original abstract.
Comparator	2	-	Basic information about comparators is correctly extracted from the <i>Objectives</i> section of the original abstract, and supplemented with information from a) the <i>Background</i> section, to expand the abbreviation <i>TF-CBT</i> to <i>trauma-focused cognitive-behavioral therapy</i> and b) the <i>Main Results</i> section, to identify the specific control conditions evaluated in the included studies, i.e., <i>wait-list</i> and <i>minimal attention</i> .
Outcome	1	missing <i>self-reported PTSD and depression symptoms</i>	Although the primary outcomes identified in the <i>Selection Criteria</i> section of the original abstract are correctly extracted, two further outcomes are mentioned in the <i>Main Results</i> section of the abstract, i.e., <i>self-reported PTSD</i> and <i>depression symptoms</i> . Although these outcomes are mentioned in the <i>Main Findings</i> section of the generated summary, they are missing from the <i>Outcomes</i> element, resulting in the assignment of the score 1.
Quality of Evidence	1	-	In the <i>Main Findings</i> section of the generated summary, evidence quality levels accompany all results for which such a level was specified in the <i>Main Results</i> section of the original abstract. Hence a score of 1 (i.e., <i>consistent with abstract</i>) was assigned.
Currency	1	-	Based on the information about the search date provided in the <i>Search Methods</i> section of the original abstract, the <i>Methodology</i> section of the generated summary correctly states that relevant studies up to February 2019 were screened. Accordingly, a value of 1 was assigned for this dimension.

Table 6: Evaluation scores, corrective actions, rationales (where appropriate) and associated discussions for the *Population*, *Intervention*, *Comparator*, *Outcome*, *Quality of Evidence* and *Currency* dimensions

Background

Present-centered therapy (PCT) is a non-trauma, manualized psychotherapy for adults with post-traumatic stress disorder (PTSD). PCT was originally designed as a treatment comparator in trials evaluating the effectiveness of trauma-focused cognitive-behavioral therapy (TF-CBT). Recent trials have indicated that PCT may be an effective treatment option for PTSD and that patients may drop out of PCT at lower rates relative to TF-CBT.

Objectives

To assess the effects of PCT for adults with PTSD. Specifically, we sought to determine whether (1) PCT is more effective in alleviating symptoms relative to control conditions, (2) PCT results in similar alleviation of symptoms compared to TF-CBT, based on an a priori minimally important differences on a semi-structured interview of PTSD symptoms, and (3) PCT is associated with lower treatment dropout as compared to TF-CBT.

Search methods

We searched the Cochrane Common Mental Disorders Controlled Trials Register, the Cochrane Library, Ovid MEDLINE, Embase, PsycINFO, PubMed, and PTSDpubs (previously called the Published International Literature on Traumatic Stress (PILOTS) database) (all years to 15 February 2019 search). We also searched the World Health Organization (WHO) trials portal (ICTRP) and ClinicalTrials.gov to identify unpublished and ongoing trials. Reference lists of included studies and relevant systematic reviews were checked. Grey literature searches were also conducted to identify dissertations and theses, clinical guidelines, and regulatory agency reports.

Selection criteria

We selected all randomized clinical trials (RCTs) that recruited adults diagnosed with PTSD to evaluate PCT compared to TF-CBT or a control condition. Both individual and group PCT modalities were included. The primary outcomes of interest included reduced PTSD severity as determined by a clinician-administered measure and treatment dropout rates.

Data collection and analysis

We complied with the Cochrane recommended standards for data screening and collection. Two review authors independently screened articles for inclusion and extracted relevant data from eligible studies, including the assessment of trial quality. Random-effects meta-analyses, subgroup analyses, and sensitivity analyses were conducted using mean differences (MD) and standardized mean differences (SMD) for continuous data or risk ratios (RR) and risk differences (RD) for dichotomous data. To conclude that PCT resulted in similar reductions in PTSD symptoms relative to TF-CBT, we required a MD of less than 10 points (to include the 95% confidence interval) on the Clinician-Administered PTSD Scale (CAPS). Five members of the review team convened to rate the quality of evidence across the primary outcomes. Any disagreements were resolved through discussion. Review authors who were investigators on any of the included trials were not involved in the qualitative or quantitative syntheses.

Main results

We included 12 studies ($n = 1837$), of which, three compared PCT to a wait-list/minimal attention (WL/MA) group and 11 compared PCT to TF-CBT. PCT was more effective than WL/MA in reducing PTSD symptom severity (SMD -0.84 , 95% CI -1.10 to -0.59 ; participants = 290; studies = 3; $I^2 = 0\%$). We assessed the quality of this evidence as moderate. The results of the non-inferiority analysis comparing PCT to TF-CBT did not support PCT non-inferiority, with the 95% confidence interval surpassing the clinically meaningful cut-off (MD 6.83, 95% CI 1.90 to 11.76; 6 studies, $n = 607$; $I^2 = 42\%$). We assessed this quality of evidence as low. CAPS differences between PCT and TF-CBT attenuated at 6-month (MD 1.59, 95% CI -0.46 to 3.63; participants = 906; studies = 6; $I^2 = 0\%$) and 12-month (MD 1.22, 95% CI -2.17 to 4.61; participants = 485; studies = 3; $I^2 = 0\%$) follow-up periods. To confirm the direction of the treatment effect using all eligible trials, we also evaluated PTSD SMD differences. These results were consistent with the primary MD outcomes, with meaningful effect size differences between PCT and TF-CBT at post-treatment (SMD 0.32, 95% CI 0.08 to 0.56; participants = 1129; studies = 9), but smaller effect size differences at six months (SMD 0.17, 95% CI 0.05 to 0.29; participants = 1339; studies = 9) and 12 months (SMD 0.17, 95% CI 0.03 to 0.31; participants = 728; studies = 5). PCT had approximately 14% lower treatment dropout rates compared to TF-CBT (RD -0.14 , 95% CI -0.18 to -0.10 ; participants = 1542; studies = 10). We assessed the quality of this evidence as moderate. There was no evidence of meaningful differences on self-reported PTSD (MD 4.50, 95% CI 3.09 to 5.90; participants = 983; studies = 7) or depression symptoms (MD 1.78, 95% CI -0.23 to 3.78; participants = 705; studies = 5) post-treatment.

Authors' conclusions

Moderate-quality evidence indicates that PCT is more effective in reducing PTSD severity compared to control conditions. Low quality of evidence did not support PCT as a non-inferior treatment compared to TF-CBT on clinician-rated post-treatment PTSD severity. The treatment effect differences between PCT and TF-CBT may attenuate over time. PCT participants drop out of treatment at lower rates relative to TF-CBT participants. Of note, all of the included studies were primarily designed to test the effectiveness of TF-CBT which may bias results away from PCT non-inferiority. The current systematic review provides the most rigorous evaluation to date to determine whether PCT is comparably as effective as TF-CBT. Findings are generally consistent with current clinical practice guidelines that suggest that PCT may be offered as a treatment for PTSD when TF-CBT is not available.

Figure 2: Original abstract from [Belsher et al. \(2019\)](#) (ID: CD012898.PUB2)

PICO Framework Summary:

Population (P): The research focused on adults diagnosed with post-traumatic stress disorder (PTSD).

Intervention (I): The primary intervention in the study was Present-centered therapy (PCT), a non-trauma, manualized psychotherapy technique designed for PTSD patients.

Comparison (C): PCT was compared to two conditions: trauma-focused cognitive-behavioral therapy (TF-CBT) and control conditions, which included wait-list or minimal attention controls.

Outcomes (O): The main outcomes measured were the severity of PTSD symptoms (based on clinician-administered measures) and treatment dropout rates.

Methodology:

The study was a systematic review and meta-analysis of randomized clinical trials (RCTs), following Cochrane standards. Researchers screened multiple databases for relevant studies up to February 2019, including grey literature and trial portals. They selected RCTs that assessed PCT versus TF-CBT or control conditions for adults with PTSD. Data analyses included random-effects meta-analyses, subgroup analyses, and sensitivity analyses.

Main Findings

- PCT showed superior effectiveness in reducing PTSD symptoms compared to wait-list/minimal attention controls with moderate-quality evidence.
- PCT was not found to be non-inferior to TF-CBT in terms of reducing PTSD symptom severity, based on a clinician-administered measure, with low-quality evidence. However, differences between PCT and TF-CBT reduced over time at 6 and 12-month follow-ups.
- PCT had a lower treatment dropout rate (approximately 14% lower) than TF-CBT, supported by moderate-quality evidence.
- No significant differences were observed in self-reported PTSD or depression symptoms post-treatment.

Conclusions

The research concluded that PCT is an effective treatment for reducing PTSD severity compared to control conditions, although it does not match TF-CBT's effectiveness according to clinician ratings. PCT's lower dropout rates suggest it may be a preferable option for patients likely to disengage from therapy. The study highlights the limitations due to the included studies being primarily designed to test TF-CBT, potentially biasing against PCT. Overall, it aligns with clinical guidelines suggesting PCT as an alternative when TF-CBT is unavailable.

Significance in Patient Care:

The research underscores PCT as an effective treatment alternative for PTSD, especially when access to TF-CBT is limited. Its lower dropout rates are clinically significant for maintaining patient engagement and adherence to therapy. However, the evidence supporting its non-inferiority to TF-CBT is of low quality, suggesting a need for further research focusing specifically on PCT's standalone efficacy.

Limitations and Recommendations for Future Research

The study acknowledges potential bias due to the primary design focus on TF-CBT effectiveness in the included trials. Future research should prioritize directly evaluating PCT's efficacy against other evidence-based treatments with distinct emphasis on PCT itself, employing high-quality RCTs to solidify its position in PTSD treatment protocols.

Figure 3: Summary of abstract from [Belsher et al. \(2019\)](#) (ID: CD012898.PUB2) generated by GPT-4o

Dimension	Score	Action	Rationale	Discussion
Methodology	1	Add info	Missing <i>number of studies</i>	Relevant methodological information is contained within multiple original abstract sections (i.e., <i>Search Methods</i> provides details of the types of resources searched; <i>Selection Criteria</i> specifies studies of interest were randomised clinical trials(RCTs); <i>Data Collection and Analysis</i> contains details about the standards used for data screening and collection, along with the data analysis methods employed; and <i>Main Results</i> contains information about the number of included studies). While most of this information was extracted and appropriately summarised, the score of 1 was assigned according to the summary's omission of a single piece of information considered essential for professional acceptability (i.e., the number of included studies).
Main Findings	2	None	-	All results reported in the <i>Main Results</i> section of the original abstract correctly extracted and appropriately summarised in the generated text, covering the comparison of the intervention of interest (i.e., <i>PCT</i>) with the target comparators (i.e., <i>TF-CBT</i> and the control conditions <i>wait-list</i> and <i>minimal attention</i>) in the context of all of the outcomes of interest (i.e., <i>PTSD severity</i> (determined by a clinician-administered measure), <i>treatment dropout rates</i> , <i>self-reported PTSD</i> and <i>depression symptoms</i>).
Conclusion	1	Remove info	<i>The summary should remove "Future research should prioritize directly evaluating PCT's efficacy against other evidence-based treatments with distinct emphasis on PCT itself, employing high-quality RCTs to solidify its position in PTSD treatment protocols."</i>	In the generated summary, concluding information is split between the <i>Conclusions</i> , <i>Significance in Patient Care</i> and <i>Limitations and Recommendations for Future Research</i> sections. Among these generated sections, the content of the <i>Conclusions</i> section is fully consistent with the information provided in the <i>Authors' conclusion</i> section of the original abstract. Appropriate content regarding implications for patient care and study limitations from the <i>Authors' conclusion</i> section of the original abstract is also used to generate accurate statements in the <i>Significance in Patient Care</i> and <i>Limitations and Recommendations for Future Research</i> sections of the summary. However, the <i>Authors' conclusion</i> section of the original abstract does not mention any potential future work. Accordingly, the final sentence in the <i>Limitations and Recommendations for Future Research</i> section of the generated summary, which recommends future work, is unsupported by the original abstract. Thus, a score of 1 is assigned.

Table 7: Evaluation scores, corrective actions, rationales (where appropriate) and associated discussions for the *Methodology*, *Main Findings* and *Conclusions* dimensions

Reasoning, Contrastive, and In-Context Strategies for Opioid Use Stage Detection on Social Media

Vinu Ekanayake and Ramakanth Kavuluru

University of Kentucky, Lexington, KY USA
{vinu.ekanayake, ramakanth.kavuluru}@uky.edu

Abstract

The opioid epidemic has ravaged the US for the past two decades and is still a persistent threat. During the same time, the increasing use of social media has created a new avenue for people to share their journeys regarding opioid use. In this context, research in automatically determining opioid use stages (e.g., misuse, addiction, recovery) based on self disclosures in social media posts is gaining traction. In this paper, using a recent benchmark, we assess different supervised strategies for identifying self-disclosed opioid use stages from Reddit posts. We consider distilled reasoning traces from DeepSeek R1 (an open weights reasoning model), supervised contrastive learning (SCL), and few-shot in-context learning (ICL) with GPT-5 to conduct a variety of experiments with encoder and encoder-decoder models. We also conduct direct zero-shot (ZS) experiments with GPT 5 and GPT 5.2. Across different models and datasets, our strategies provide improvements in performance with some nuances that are too subtle to elaborate in the abstract. A surprising finding is that ZS results with GPT-5 are better than all supervised results, which ushers a new frontier for LLM-based classification of opioid use in social media posts. Our code is available for reuse and replication: <https://github.com/bionlproc/Opioid-Stage>.

Keywords: Opioid use stages, contrastive learning, test time reasoning, zero shot learning

1. Introduction

The opioid crisis remains a major public health concern in the United States. It has ravaged America for over two decades with drug overdose deaths reaching a peak of over 110K in 2022, 76% of which are attributed to opioids (Spencer et al., 2024). Opioid use disorder (OUD) is a related chronic condition characterized by repeated use in the past 12 months. Specifically, OUD is defined based on a variety of DSM criteria (National Center for Injury Prevention and Control, 2024) that are further converted to ICD-10 codes. OUD diagnosis facilitates identification, tracking, and intervention for individuals whose situation may exacerbate to overdoses and more serious long term disability. With regards to both overdoses and OUD, the situation appears to have stabilized since peak levels during the Covid-19 pandemic, as per the latest report from the US Substance Abuse and Mental Health Services Administration (2025).

Treatment for OUD often involves both behavioral and pharmaceutical strategies (Wakeman et al., 2020) and both aspects are discussed by users on social media at length, especially on Reddit, which has emerged as a key platform for peer support and recovery discussions (Laud et al., 2025; Carpenter et al., 2025). Reddit has several dedicated forums (e.g., r/OpiatesRecovery) where individuals openly share experiences and seek advice. Complementing the clinical definitions of OUD, researchers have identified discrete stages of opioid use to characterize the journey of a user, warranted by the need to design inter-stage transition-specific interven-

tions (Park et al., 2020). Yang et al. (2024) used this need to come up with six stages: Not Using, Medical Use, Misuse, Addiction, Recovery, and Relapse, following the definitions by Smith et al. (2013) and National Institute on Drug Abuse (2007).

In their seminal effort, Yang et al. (2024) created a dataset and conducted baseline experiments to classify Reddit posts discussing opioids into one of these six stages. Our current paper deals with how we can improve over their baseline experiments. At this point, we note that their dataset included a so called “explanation” for each input Reddit post indicating the spans in the input that highlight the textual evidence for the annotator’s decision to assign a stage. This explanation was used as part of the input to the model (along with the original post). Since these explanations are not available at test time, they designed a strategy to extract a “silver” explanation that is generated by a different supervised model. The silver explanation, in this sense, does not need human inputs at test time. However, our methods based on recent advances in reasoning LLMs obviate this need for these silver explanations. As such, for brevity and space limitations, we do not discuss or use the silver explanations in all reported results of this paper. Due to scarcity of expert time to create training labels, Yang et al. (2024) used Amazon Mechanical Turk to create significant portions of the dataset. The entire training dataset consists of these so called “novice” labeled examples while the test set has both expert and novice labeled instances.

Some of the main challenges on this task is the lack of expert labeled training examples, nontrivial

semantic overlap between the stages, and limited reasoning fidelity (and hence interpretability) of encoder and encoder-decoder models typically used in classification tasks. We next outline how we address these aspects in this paper.

Accurate stage detection often requires nuanced logical deduction beyond surface-level cues. Recent reasoning-capable LLMs like DeepSeek-R1 can produce detailed chain-of-thought traces or summarized reasoning at inference time, and we already have evidence that bigger more powerful models can improve downstream model performance when distilled into smaller models (Wei et al., 2022; Hsieh et al., 2023). However, it remains unclear whether such reasoning is necessary for opioid stage classification. We use DeepSeek R1 as a teacher model to generate two forms of reasoning: (a) *summarized reasoning* – concise 2–3 sentence justifications highlighting the most salient linguistic cues, and (b) *step-by-step reasoning* – detailed internal reasoning traces capturing the model’s multi-step deductions. By distilling these into smaller student models (T5-3B, T5-11B, and DeBERTa v3), we test whether explicit reasoning can improve classification accuracy.

Furthermore, Yang et al. (2024) identified that opioid use stage classification suffers from significant semantic overlap, particularly between adjacent stages like Misuse and Addiction or Recovery and Relapse, as illustrated in Table 1. To address this, we integrate supervised contrastive learning (SCL), which, as shown by Gunel et al. (2020), enhances inter-class separability by explicitly pushing apart representations of different stages in the latent space. Finally, given the scarcity of expert labels in existing opioid use corpora and the persistent discrepancy between expert and novice annotations, we reduce our reliance on noisy novice data by employing in-context learning (ICL). We use GPT-5 as a synthetic annotator grounded in a small set of expert examples to realign novice-labeled training data toward expert standards. By systematically evaluating these three strategies, our work provides a comparison of how reasoning-centric, data-centric, and representation-centric techniques impact classification across different model scales.

Our results show that the optimal reasoning format depends on the evaluation setting: summarized reasoning yields the strongest performance on the expert test set, while step-by-step reasoning achieves the best results on the novice test set. The integration of SCL improves performance by enhancing inter-class discrimination, while ICL-based realignment produces higher agreement with expert test sets. While none of our methods is a silver bullet, these techniques yield nontrivial gains in both accuracy and interpretability, demonstrating a promising pathway toward scalable, expert-

aligned opioid use monitoring from social media. The dataset is only available from its creators (Yang et al., 2024) but all our code is available for reuse and replication: <https://github.com/bionlproc/Opioid-Stage>.

2. Related Works

As already discussed, Yang et al. (2024) demonstrated that explanation-augmented models improved detection of opioid use stages from Reddit posts. Earlier work examined substance-use detection more broadly from forums and Twitter, in identifying misuse, overdose risk, and relapse signals from user-generated content (Fodeh et al., 2021; Raza et al., 2023; Ahmad et al., 2026; Yang et al., 2022). Our lab has recently showed how Reddit posts can be used to glean evolving barriers to opioid recovery (Ekanayake et al., 2025).

Prior work has shown that models can benefit from human or model-generated rationales, which enhance both predictive accuracy and interpretability. Camburu et al. (2018) and Rajani et al. (2019) demonstrated that jointly training models to generate explanations alongside labels yields superior performance in reading comprehension and commonsense reasoning. Recent reasoning LLMs such as DeepSeek R1 (Guo et al., 2025) and chain-of-thought (COT) prompting have shown that decomposing complex problems into intermediate, step-by-step reasoning steps significantly improves robustness and logic in large language models (Wei et al., 2022). Subsequent work has further demonstrated that these reasoning rationales can be distilled to train smaller, more efficient, and effective task-specific models (Hsieh et al., 2023; Hancock et al., 2018). Our work applies these ideas to a high-stakes, fine-grained social health text task and compares the effectiveness of summarized vs step-by-step rationales.

Supervised contrastive learning has been shown to improve representation quality in classification problems by explicitly pulling together representations of same-label examples and pushing apart different labels (Khosla et al., 2020; Gunel et al., 2020; Kim et al., 2022). In health NLP, contrastive objectives have been used to better separate semantically adjacent classes, for example in automated diagnostic coding from clinical notes (Kailas et al., 2023). Separately, in-context learning and related prompting strategies with powerful LLMs have been used for expert-like few-shot annotation and label refinement in clinical and mental-health tasks, including clinical NER, suicidality phenotyping, and OUD identification from electronic health records (Baroian, 2025; Li et al., 2025; Huang et al., 2025; Molina et al., 2025). We build on this line of work by using GPT-5 to realign noisy novice labels

OOD Stage	Reddit Post Excerpt (Title + Text)
Medical Use	[TITLE] Codeine doses [TEXT] So I was recently prescribed Tylenol 3's with 30mg Codeine. Is it safe to take two, ingesting 60mg of codeine? Has anyone ever done it with this amount or more? Comment below!
Misuse	[TITLE] I'm on tramadol first time [TEXT] I took 350mg first time on this, it's kinda underwhelming but it's lght. I hit a juul and that took me way higher but idk I expected more from this
Addiction	[TITLE] Hydrocodone dosage [TEXT] Is 30mg too much if I just started taking it daily a week ago? I can tell my tolerance is going up so I'm just wondering if I can start to slowly increase my dosage.
Recovery	[TITLE] Venting [TEXT] I'm just ready to feel like myself again I won't give up though My life is not so bad but my mind is hopefully it a be over soon ...stay strong everybody it does get easier
Relapse	[TITLE] Withdrawal question please help? [TEXT] So I tried to get clean went 4 days and I was starting to get over the withdrawal then I used again will I feel the effects of the withdrawal all over again?

Table 1: Representative opioid use stage examples from the training dataset

toward a small expert subset and then measuring how this affects downstream models.

3. Methodology

Recall our goal is to improve the fine-grained classification of Reddit posts into six opioid use stages (*Not Using*, *Medical Use*, *Misuse*, *Addiction*, *Recovery*, and *Relapse*) using the dataset from Yang et al. (2024). The dataset is created using opioid related posts from three subreddits (*r/OpiatesRecovery*, *r/Opiates*, and *r/Drugs*) with sufficient accommodations to only retain opioid related posts. The inputs to the models consist of the Reddit post's title concatenated with its body text. To ensure comparisons with prior work, we followed the exact data partitioning from Yang et al. (2024):

- **Novice-annotated data:** 1936 training instances and 150 test instances,
- **Expert-annotated data:** 442 test instances (experts do not contribute to training).

All models are evaluated across the two test sets (novice and expert) to align with the evaluation framework introduced in Yang et al. (2024). The top class in novice annotated posts is *addiction* (29%) but for expert dataset it is *not using* (34%). To illustrate the nature of the self-disclosed Reddit posts and the fine-grained opioid use stages, we provide representative examples from the dataset introduced by Yang et al. (2024) in Table 1. These examples demonstrate how user narratives are categorized into distinct phases, highlighting the contextual nuances our models aim to capture.

3.1. Reasoning Distillation

Reasoning distillation aims to improve downstream model performance by providing student models with explicit intermediate rationales. These rationales reflect how a stronger teacher model justifies its predictions. To generate them, we use DeepSeek R1, a recent large language model optimized for structured reasoning (Guo et al., 2025). For each training instance, DeepSeek R1 is prompted in a zero-shot setting with the Reddit post and its ground-truth stage label, and asked to produce an explanation for why the post corresponds to that label. The full prompting template used to obtain reasoning is provided in Figure 1 of the Appendix.

3.1.1. Reasoning Formats

We generate two forms of teacher rationales:

- *Summarized reasoning:* A concise explanation (with 2-3 sentences) in which DeepSeek R1 provides a high-level justification for the label, highlighting the most salient cues or linguistic indicators present in the post. These summaries average approximately 62 words in length.
- *Step-by-step reasoning:* A detailed internal reasoning trace obtained from the model's `<think>` segment. These traces represent DeepSeek R1's hidden intermediate deliberations, including multi-step deductions, latent evidence extraction, and fine-grained decision heuristics. Compared to the summarized rationales, these internal reasoning sequences are significantly longer and more granular, av-

eraging approximately 322 words per post, reflecting the full chain-of-thought the model produces before emitting its final answer.

The two formats allow us to evaluate whether student models benefit more from concise rationales or from richer, model-internal reasoning traces.

3.1.2. Student Model Training

The reasoning outputs are incorporated directly into model supervision for T5-3B and T5-11B models (Raffel et al., 2020). During fine-tuning, the student models are trained to generate both the label y_i followed by its associated reasoning r_i (summarized or step-by-step) given input x_i , using a conditional language modeling objective:

$$\mathcal{L}_{\text{CE}} = - \sum_i \log P_{\theta}(y_i, r_i \mid x_i).$$

similar to other efforts (Wadhwa et al., 2023). This setup enables us to quantify the impact of distilled reasoning on classification accuracy and to compare the relative benefits of short versus detailed reasoning representations.

3.2. Supervised Contrastive Learning

To improve the separability of closely related opioid use stages, we incorporate supervised contrastive learning (SCL) as a two-stage training procedure applied uniformly across all model architectures evaluated in this work.

3.2.1. Stage 1: Contrastive Pretraining

In the first stage, the model encoder is optimized using a supervised contrastive objective. Each training batch is constructed using a balanced class sampling strategy, where a fixed number of examples are drawn from each of the 6 opioid use stages. Two examples per class are used in all models. This construction ensures that all batches contain well-defined positive and negative relationships.

View generation For each text instance x_i , we generate a single augmented variant $\tilde{x}_i$ using a random text-level operation (synonym substitution, token deletion, or token swapping). Both the original and augmented inputs are encoded independently:

$$v_i^{(1)} = f_{\theta}(x_i), \quad v_i^{(2)} = f_{\theta}(\tilde{x}_i),$$

where $f_{\theta}(\cdot)$ denotes the encoder.

Projection and normalization The encoded representations are passed through a two-layer projection head (Linear $\rightarrow$ ReLU $\rightarrow$ Linear $\rightarrow$ Layer-Norm). We apply L2 normalization to the resulting projection to produce unit vectors (z_i), suitable for contrastive learning.

Contrastive objective We adopt the supervised contrastive loss from Khosla et al. (2020). Let $I = \{1, \dots, 2N\}$ be the set of indices for all augmented views in the batch. For an anchor embedding $i \in I$, let y_i be its corresponding opioid stage label. The set of positive indices is defined as

$$P(i) = \{p \in I \mid y_p = y_i \text{ and } p \neq i\},$$

representing all other views in the batch with the same label, while $A(i) = I \setminus \{i\}$ contains all remaining views. The loss is then defined as:

$$L = \sum_{i=1}^{2N} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(z_i, z_p)/\tau)}{\sum_{a \in A(i)} \exp(\text{sim}(z_i, z_a)/\tau)},$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity computed using the dot product between L2-normalized embeddings, and $\tau = 0.07$ is the temperature parameter. This step yields an encoder whose latent space is structured to better reflect stage distinctions.

3.2.2. Stage 2: Task-specific Fine-tuning

Following contrastive pretraining, we load the resulting encoder weights into the downstream opioid stage classifier. To preserve the structure learned during SCL and mitigate catastrophic forgetting, we freeze the lower portion of the encoder and fine-tune only the upper k layers (determined based on validation experiments) alongside the prediction head (or decoder in seq2seq settings).

3.3. Zero-shot with GPT5

To establish a strong baseline we evaluated the most recent OpenAI large language models (GPT-5 and GPT-5.2) in a zero-shot setting, that is, without any task-specific fine-tuning or in-context examples. Two prompting strategies were explored: (1) Simple prompt (baseline), which directly mirrors the prompt used in the original study for zero-shot inference (Figure 3 in the Appendix). This minimal formulation tests the model’s ability to map raw Reddit text to the six opioid use stages with only the label list as guidance. (2) Guideline-enhanced prompt, that incorporates the full class definitions that were provided to the human annotators in the original dataset (Figure 4 in the Appendix). Both prompts were submitted to GPT-5 and GPT-5.2 via the OpenAI API with default temperature (0) and greedy decoding to obtain a deterministic label.

3.4. Novice-to-Expert Realignment of Training Data

While reasoning distillation and supervised contrastive learning improve model learning dynamics and representation quality, their effectiveness remains ultimately bounded by the quality of the

underlying training labels. Note that the training split is exclusively novice-only; as per Yang et al. (2024) the novice turkers need only a 60% accuracy threshold, on a small sample of expert labeled instances, to be selected for this task. Thus testing on expert annotated test data is expected to give worse performance. To address this, we use few-shot in-context learning (ICL) with GPT-5 (OpenAI, 2025) to realign novice labels toward expert annotation standards.

We begin by selecting 40 expert-labeled instances, with 5-8 examples per class depending on availability. This selection counters the strong class imbalance in the expert subset, (e.g., Not Using: 34.51% vs. Medical Use: 3.52%), while ensuring that each stage is represented by multiple high-quality exemplars. From the pool of 40 expert examples, we construct 10 distinct ICL prompt sets, chosen to balance diversity of expert demonstrations with computational feasibility. Each set contains one demonstration per class (six total), providing GPT-5 with a balanced snapshot of expert decision patterns across all stages.

For each novice-labeled post, GPT-5 is prompted ten times, once with each of the ten ICL demonstration sets. Each prompt contains the full annotation guidelines used by Yang et al. (2024), six expert few-shot exemplars, and the target Reddit post. GPT-5 is instructed to produce a JSON-formatted output containing a corrected label selected from the six opioid use stages and a rationale consisting of the minimal text span from the post that justifies the prediction. This process yields ten independent candidate labels for each instance. The final “aligned” label is determined through majority voting over these ten predictions. The complete prompt template is provided in Figure 2 of the Appendix.

The resulting relabeled dataset is used to train the same models evaluated in the main experiments, including both label-only and reasoning-augmented variants. We refer to this curated version as *ICL-aligned*. To avoid any form of data leakage, the 40 expert examples selected for the ICL demonstration sets were excluded entirely from both training and evaluation, and were removed from the original expert test split so that no item used in the ICL prompts appeared in the test set.

4. Results

We evaluate our framework across three dimensions: (1) reasoning distillation, comparing summarized vs step-by-step rationales; (2) supervised contrastive learning (SCL); and (3) expert realignment via ICL. All models were trained on novice-labeled dataset, and evaluated on both the novice and expert test sets using accuracy and macro F1 score. Each method is compared against baseline models that we reproduced (as described in Section 4.1)

using the original input–output format (title + post → label), so that observed improvements can be attributed to methodological changes rather than optimization differences.

4.1. Reproduced Baselines

To enable fair comparison across all methods evaluated in this study, we first reproduced the baseline models from the original paper by Yang et al. (2024) using the same input-output with updated training configurations (see Table A in the Appendix for full hyperparameter specifications across all models). Table 4.2.1 reports the performance for all architectures, for both test sets. Most models showed modest improvements of 1-3 points over the originally reported numbers (Appendix’s Table A), with T5-3B exhibiting the largest improvements on both expert and novice splits with improvements of 4-6 points. We interpret all reasoning distillation, SCL, and ICL results with respect to these reproduced baselines.

4.2. Reasoning Distillation Results

Our initial experiments evaluated the effectiveness of reasoning distillation using two distinct reasoning formats generated by DeepSeek R1: *summarized* (Table 4.2.1) and *step-by-step traces* (Table 4.2.1), evaluated for T5-3B and T5-11B.

4.2.1. Comparison with Reproduced Baselines

Relative to the reproduced baselines, reasoning distillation has a clear positive effect on the smaller T5-3B model and a consistently negative effect on T5-11B. For T5-3B, adding reasoning supervision improves performance on novice test set: step-by-step rationales increase novice F1 from 0.755 in the baseline to 0.791, an improvement of 3.6 F1 points, while summarized reasoning also improves novice F1 to 0.784. On the expert test set, both reasoning variants slightly reduce F1 relative to the baseline (0.703), producing 0.698 (step-by-step) and 0.683 (summarized), although accuracy increases marginally in both cases (1-2 percentage points). For the T5-11B model, reasoning distillation consistently degrades performance over standard fine-tuning. On the expert test set, F1 drops from 0.739 to 0.723 (summarized) and 0.674 (step-by-step), and on the novice test set from 0.814 to 0.763 and 0.809, respectively. These results indicate that, at this scale, additional reasoning supervision does not provide clear benefits and can interfere with the model’s pre-trained decision boundaries.

4.2.2. Impact of Reasoning Format

The comparison between summarized and step-by-step formats reveals that the optimal reasoning

Model	Dataset	Accuracy	F1
DeBERTa-v3-Base	Expert	0.681	0.675
	Novice	0.707	0.712
DeBERTa-v3-Large	Expert	0.708	0.707
	Novice	0.727	0.737
T5-3B	Expert	0.690	0.703
	Novice	0.747	0.755
T5-11B	Expert	0.732	0.739
	Novice	0.813	0.814

Table 2: Reproduced baseline results for DeBERTa and T5 Models

Model	Dataset	Accuracy	F1
T5-3B	Expert	0.713	0.683
	Novice	0.787	0.784
T5-11B	Expert	0.733	0.723
	Novice	0.760	0.763

Table 3: Models trained with summarized reasoning

Model	Dataset	Accuracy	F1
T5-3B	Expert	0.708	0.698
	Novice	0.787	0.791
T5-11B	Expert	0.683	0.674
	Novice	0.807	0.809

Table 4: Models trained with step-by-step reasoning

format depends on student model size and dataset. For T5-3B, step-by-step reasoning is consistently more effective than summarized reasoning on both expert and novice test sets. On expert annotations, step-by-step achieves higher F1 than summarized (0.698 vs 0.683, +1.5 points), and on novice annotations it achieves the best overall performance with F1 0.791 compared to 0.784 for summarized reasoning. For T5-11B, step-by-step reasoning substantially outperforms summarized reasoning on the novice test set (F1 0.809 vs. 0.763), whereas summarized reasoning is less harmful on the expert set (F1 0.723 vs. 0.674), though both formats remain below the baseline. Taken together, these results show that explicit reasoning supervision is most helpful for the smaller T5-3B model, especially on the noisier novice data where step-by-step traces produce the largest gains over the baseline. For the larger T5-11B model, which already performs strongly under standard fine-tuning, reasoning distillation reduces performance on both expert and novice test sets, suggesting that reasoning supervision may be ill-suited for larger models.

4.3. Supervised Contrastive Learning

We next evaluate the impact of integrating an SCL objective on all our model architectures (Table 4.3). Each SCL-enhanced model is directly comparable to its baseline counterpart in Table 4.2.1.

SCL generally improves performance on novice-labeled test sets for smaller and mid-sized models, while its effect on very large models is mixed. For DeBERTa-v3-Base, DeBERTa-v3-Large, and T5-3B, SCL yields substantial F1 gains on novice data, whereas T5-11B experiences slight degradation in this regime. On expert test set, the impact of SCL is more modest and model-dependent.

On the expert test set, SCL produces small but noticeable changes. For T5-3B, SCL improves expert F1 from 0.703 to 0.724 (+2.1 points), alongside a corresponding 2 point increase in accuracy. DeBERTa-v3-Base and DeBERTa-v3-Large show only modest shifts in expert F1 under SCL, remaining close to their respective baselines. In contrast, T5-11B experiences a clear degradation: expert F1 decreases from 0.739 to 0.690 (-5 points), and accuracy falls from 0.732 to 0.686.

On the novice test set, the impact of SCL is considerably larger particularly for smaller and mid-sized models. For T5-3B, SCL increases novice F1 from 0.755 to 0.820 (+6.5 points), with accuracy rising from 0.747 to 0.800. DeBERTa-v3-Base and DeBERTa-v3-Large also achieve strong novice performance under SCL, with F1 scores of 0.726 and 0.756, respectively indicating robustness to noisy labels compared to their non-SCL baselines. By contrast, T5-11B exhibits a slight reduction in novice F1, from 0.814 to 0.803, suggesting that SCL is less beneficial at this scale in the realistic inference setting.

These patterns suggest that SCL is especially beneficial for DeBERTa-v3-Base and T5-3B under noisy novice supervision, where it yields multi-point F1 gains and more stable performance. Larger models such as DeBERTa-v3-Large and T5-11B exhibit more limited benefits.

Model	Dataset	Accuracy	F1
DeBERTa-v3-Base	Expert	0.677	0.660
	Novice	0.720	0.726
DeBERTa-v3-Large	Expert	0.710	0.703
	Novice	0.753	0.756
T5-3B	Expert	0.710	0.724
	Novice	0.800	0.820
T5-11B	Expert	0.686	0.690
	Novice	0.800	0.803

Table 5: Effect of Supervised Contrastive Learning (SCL) on model performance.

Model	Dataset	Prompt w/o guidelines		Prompt w/ guidelines	
		Accuracy	F1	Accuracy	F1
GPT-5	Expert	0.649	0.647	0.79	0.796
	Novice	0.625	0.593	0.749	0.739
GPT-5.2	Expert	0.652	0.644	0.739	0.722
	Novice	0.622	0.588	0.749	0.739

Table 6: Zero-shot performance of GPT-5 models with/without guidelines, averaged across 3 runs.

4.4. Zero-shot Performance with GPT-5

Table 4.3 lists zero-shot performances of GPT-5 and GPT-5.2 on both expert and novice test sets. With the concise prompt from the original paper, GPT-5 attained an accuracy of 0.649 and an F1 score of 0.647 on the expert set, and 0.625 accuracy (F1 0.593) on the novice set. GPT-5.2 performed similarly on both sets, achieving 0.652 accuracy (F1 0.644) for expert data and 0.622 accuracy (F1 0.588) for novice data. But introducing the elaborate guideline-enhanced prompt consistently improved performance for both GPT-5 and GPT-5.2 across both expert and novice datasets. GPT-5 achieved an F1 score of 0.796 on the expert dataset, a 15 point bump compared to the simple prompt. Notably, for GPT-5, the performance on the expert-labeled data was higher than on the novice-labeled data, even surpassing the results of the newer GPT-5.2 by 7.4 points (F1 0.722 vs 0.796). This comparison suggests that for zero-shot stage classification, GPT-5 demonstrated stronger performance on expert data with the enhanced prompting strategy compared to its successor.

4.5. Realignment via ICL

Our final set of experiments assessed whether few-shot ICL with GPT-5 could realign noisy novice-labeled data toward expert annotation standards and hence produce better performance on the expert test set.

On the expert test set, ICL-based realignment gave mixed but often positive effects across all architectures, particularly in accuracy (Table 4.5). For

DeBERTa-v3-Base, expert F1 increases from 0.675 in the baseline to 0.694, a gain of around 2 points. T5-3B also shows a modest improvement in expert F1, rising from 0.703 to 0.718 (+1.5 points), with accuracy increasing from 0.690 to 0.744. Conversely, DeBERTa-v3-Large experiences a slight decline in expert F1, from 0.707 to 0.682 (-2.5 points). However, the accuracy rises by 2 points. T5-11B also sees a nearly 2 point decrease in F1 score, with the accuracy slightly increasing from 0.732 to 0.736.

However, the improvements in the expert-set performance were accompanied by substantial declines in novice test set performance. Most notably, T5-3B’s accuracy dropped from 0.840 to 0.727 after ICL realignment (-11.3 points). So, ICL-based label refinement successfully achieves its stated objective, improved alignment with expert evaluation standards using only 40 expertly annotated exemplars. This efficiency is operationally valuable for scenarios where expert evaluation is the primary validation metric. The reduction in performance on the novice evaluation set is expected, as the realigned labels impose stricter expert-style decision boundaries that differ from the noisier patterns present in novice annotations.

5. Discussion

Our work addresses three persistent challenges in fine-grained opioid use stage classification from social media data: (1) limited high-quality expert annotations, (2) semantic overlap across stages, and (3) the lack of interpretable, reasoning-driven models suitable for public health applications. Our find-

Model	Dataset	Accuracy	F1
DeBERTa-v3-Base	Expert	0.734	0.694
	Novice	0.647	0.620
DeBERTa-v3-Large	Expert	0.729	0.682
	Novice	0.693	0.663
T5-3B	Expert	0.744	0.718
	Novice	0.747	0.739
T5-11B	Expert	0.736	0.720
	Novice	0.713	0.697

Table 7: Performance of models trained on ICL-realigned data.

ings showed that reasoning distillation, supervised contrastive learning, and ICL-based label realignment, each address different challenges in stage classification, with distinct benefits depending on model size and label quality.

5.1. Reasoning Distillation

The reasoning distillation experiments highlight the importance of explanation format in transferring knowledge from a larger teacher model into smaller student architectures. We find that for smaller models such as T5-3B, both summarized and step-by-step explanations provide performance gains, but their effectiveness varies by explanation quality. Summarized reasoning proves more helpful on the expert test set, while step-by-step reasoning provides the largest improvements on the noisier novice test set. We hypothesize that explicit reasoning steps help smaller models disentangle ambiguous decision boundaries by exposing the underlying decision-making process, which is valuable when training labels are noisy. These patterns suggest that detailed reasoning helps a smaller model handle noisy novice labels by exposing underlying decision patterns, whereas concise rationales better align with expert decision boundaries.

However, reasoning distillation does not uniformly benefit larger models. For T5-11B, both summarized and step-by-step explanations tend to degrade performance relative to standard fine-tuning. Summarized reasoning produces only marginal changes on the expert set and step-by-step reasoning can substantially reduce expert accuracy. Since larger models such as T5-11B already learn to infer implicit reasoning patterns during standard fine-tuning, adding explicit teacher rationales may over-constrain their learned representations or introduce conflicting supervision signals, leading to performance degradation. This trade-off suggests that reasoning augmentation is most effective for smaller models on noisy data, where the explicit signal provides genuine benefit.

To illustrate the impact of reasoning-aware supervision, we present a representative example in

Figure 5 in the Appendix, where the baseline fine-tuned model fails to predict the correct stage, while the reasoning-distilled T5-3B model succeeds. In this instance, the baseline model misclassifies the post as Addiction, likely due to strong lexical associations with terms such as “morphine”, “high dose”, and “high”, which are frequently correlated with the Addiction class. This behavior indicates a reliance on surface-level cues rather than a deep understanding of the user’s behavioral context. In contrast, the model trained with summarized reasoning correctly predicts Misuse. Using rationales distilled from the DeepSeek-R1 teacher model, the student model may have internalized the logical distinctions required for this task. The rationale generated at inference time demonstrates that the model prioritizes key factors, such as non-therapeutic intent (seeking a “good buzz”) and limited usage frequency, rather than over-weighting the presence of high-risk keywords. This example highlights how reasoning distillation enables the model to capture subtle distinctions between semantically overlapping classes. In particular, it shows that the approach aligns the model’s decision boundaries with annotation guidelines by distinguishing intent-driven, non-therapeutic use (*Misuse*) from compulsive or dependence-driven behavior (*Addiction*).

5.2. Supervised Contrastive Learning

SCL consistently strengthened the discriminative capabilities of many models, with the largest benefits observed for smaller to mid-sized architectures particularly on the noisy novice test set. Both DeBERTa-v3-Base and DeBERTa-v3-Large exhibit clear novice F1 gains in this setting, indicating that SCL effectively helps untangle closely related opioid use stages such as Misuse versus Addiction or Recovery versus Relapse which often exhibit overlapping linguistic cues. T5-3B sees the largest relative improvement on novice data, substantially closing the performance gap to the T5-11B baseline. In fact, with SCL, T5-3B slightly surpasses the performance of the fine-tuned T5-11B baseline. On the expert set, SCL produced more modest and model-

dependent effects; T5-3B and DeBERTa-v3-Base showed small F1 gains, whereas DeBERTa-v3-Large was largely unchanged and very large models (T5-11B) showed a clear performance degradation. Overall, these results indicate that SCL acts as a noise-robust regularizer that is particularly helpful for moderate-capacity models trained on noisy novice annotations. The benefits tend to diminish or even reverse for very large models whose baselines are already strong, especially on expert-labeled data.

5.3. Zero-shot Performance with GPT-5

GPT-5 had substantial gains (+15 points in F1 on expert test set) with the detailed prompt containing full guidelines provided to human annotators compared to a brief prompt just listing the class names. It appears that current frontier LLMs equipped with reasoning capabilities are capable of digesting longer instructions and leverage them in excelling at off-the-shelf zero-shot tasks. More importantly, the expert set zero-shot F1 of 0.796 by GPT-5 is almost six points above the best supervised F1 of 0.739 by T5-11B. The GPT-5 score was based on average from three different runs and hence it is unlikely it's a fluke. Also, we do not suspect any label leakage from the dataset into GPT-5's training data considering the dataset creators did not publicly release the dataset and we could not find any other online availability based on Web search. Thus, we believe this could be a new frontier in the abilities of frontier LLMs for social media based analytics.

On the flip side, GPT-5 obtained considerable gains compared against the newer GPT-5.2 model. This outcome suggests that newer LLM versions do not always guarantee improved reliability or accuracy across all specialized applications. Without insight into proprietary training data or architectural changes, explaining these performance differences remains challenging. This highlights that newer LLM iterations may not consistently enhance or replicate the performance of prior versions on specific tasks.

5.4. Label Realignment via ICL

The ICL realignment experiments used GPT-5 as an offline oracle to shift novice-labeled training data towards expert annotation standards. This is operationally valuable because it avoids the need for extensive expert annotation to improve label quality. The relabeled training data improved expert-set performance for smaller models, specifically DeBERTa-v3-Base and T5-3B. However, it produced less consistent improvements for larger models. T5-11B, for instance, consistently showed lower F1 compared to its already strong baseline. As expected, models trained on the realigned labels showed reduced performance on the novice

evaluation set; F1 scores dropped across architectures with significant drops in the T5 models. This reflects the substantive differences between novice and expert annotation practices, despite being provided the same guidelines.

6. Conclusion

This paper provides a comprehensive evaluation of reasoning distillation, supervised contrastive learning, and expert-guided ICL realignment for fine-grained opioid use stage classification using a new benchmark dataset introduced in 2024. This is the first effort to attempt and improve upon the results from the original paper by [Yang et al. \(2024\)](#). Our results show that each method offers distinct benefits, improving interpretability, latent separability, or expert alignment depending on model size and the quality of supervision. In the end, none of our methods is a silver bullet offering gains across the board for all model sizes. Surprisingly, zero shot classification with the GPT-5 model (which already has inbuilt reasoning) with the full annotation guidelines as part of the prompt improves over best supervised results. Ultimately, method selection should be driven by the target evaluation, infrastructure, and operational constraints (e.g., privacy leakage concerns might rule out API calls to frontier LLMs in the healthcare domain). Together, our findings offer practical ideas for building more reliable, interpretable, and expert-aligned public-health NLP systems under realistic constraints.

7. Limitations

Several limitations warrant consideration. First, reasoning distillation depends on the quality of teacher explanations; alternative teacher models or more structured prompting strategies may yield further improvements. Second, while SCL improves class separability for many models, its interaction with larger models (e.g., T5-11B) remains less clear, with diminishing returns on expert data. Third, the ICL realignment process depends entirely on the quality of the teacher model used. Moreover, one methodological consequence of our expert realignment procedure is that the 40 expert-annotated examples used to construct the ICL demonstration sets were removed from the original expert test split to prevent data leakage. As a result, our expert evaluation set is not identical to that used in [Yang et al. \(2024\)](#), and direct numerical comparison to their expert-test results should be interpreted with caution (just for the ICL results in Table 4.5). With all our strategies, while improvements on smaller models are nice, the gains are simply not there or minimal with larger models such as T5-11B. What would it take to improve larger models would be part of our future work.

8. Acknowledgments

This work is supported by the U.S. National Institute on Drug Abuse through grant R01DA057686. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

9. Bibliographical References

- Muhammad Ahmad, Rita Orji, Maaz Amjad, et al. 2026. Automated risk assessment of opioid use: Analysis using pre-trained transformers on social media data. *JMIR infodemiology*, 6:e77783.
- Andrei Baroian. 2025. Supervised fine-tuning or in-context learning? evaluating llms for clinical ner. *arXiv preprint arXiv:2510.22285*.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, et al. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.
- Kristy A Carpenter, Anna T Nguyen, Delaney A Smith, et al. 2025. Which social media platforms facilitate monitoring the opioid crisis? *PLOS Digital Health*, 4(4):e0000842.
- 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.
- Samah Jamal Fodeh, Mohammed Al-Garadi, Osama Elsankary, et al. 2021. Utilizing a multi-class classification approach to detect therapeutic and recreational misuse of opioids on twitter. *Computers in biology and medicine*, 129:104132.
- Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.
- Daya Guo, Dejian Yang, Haowei Zhang, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Braden Hancock, Paroma Varma, Stephanie Wang, et al. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895.
- Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, et al. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.
- Ming Huang, Zehan Li, Yan Hu, et al. 2025. Multi-label classification with generative ai models in healthcare: A case study of suicidality and risk factors. *ArXiv*, pages arXiv–2507.
- Prajwal Kailas, Max Homilius, Rahul C Deo, et al. 2023. Notecontrast: Contrastive language-diagnostic pretraining for medical text. In *Machine Learning for Health (ML4H)*, pages 201–216. PMLR.
- Prannay Khosla, Piotr Teterwak, Chen Wang, et al. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.
- Jaewon Kim, Hyukjong Lee, Jooyoung Chang, et al. 2022. Generalized supervised contrastive learning. *arXiv preprint arXiv:2206.00384*.
- Tanmay Laud, Akadia Kacha-Ochana, Steven A Sumner, et al. 2025. Large-scale analysis of online questions related to opioid use disorder on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1068–1084.
- Zehan Li, Wanjing Wang, Lokesh Shahani, et al. 2025. Explainable suicide phenotyping from initial psychiatric evaluation notes using reasoning large language models. *medRxiv*, page 03.
- Melanie Molina, Cynthia Fenton, Kathy T Le-Saint, et al. 2025. Comparing computable structured phenotype-versus large language model-identification of opioid use disorder using electronic health record data. *medRxiv*, page 12.
- National Center for Injury Prevention and Control. 2024. Opioid Use Disorder: Diagnosis. <https://www.cdc.gov/overdose-prevention/hcp/clinical-care/opioid-use-disorder-diagnosis.html>. Centers for Disease Control and Prevention. Accessed: 2026-02-25.
- National Institute on Drug Abuse. 2007. *Drugs, Brains, and Behavior: The Science of Addiction*. National Institute on Drug Abuse, National Institutes of Health, U.S. Department of Health and Human Services. NIH Publication.
- OpenAI. 2025. [Introducing gpt-5](#).
- Ju Nyeong Park, Saba Rouhani, LEO Beletsky, et al. 2020. Situating the continuum of overdose

- risk in the social determinants of health: a new conceptual framework. *The Milbank Quarterly*, 98(3):700–746.
- Colin Raffel, Noam Shazeer, Adam Roberts, et al. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, et al. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4932–4942.
- Shaina Raza, Brian Schwartz, Sahithi Lakamana, et al. 2023. A framework for multi-faceted content analysis of social media chatter regarding non-medical use of prescription medications. *BMC digital health*, 1(1):29.
- Shannon M Smith, Richard C Dart, Nathaniel P Katz, et al. 2013. Classification and definition of misuse, abuse, and related events in clinical trials: Acttion systematic review and recommendations. *Pain@*, 154(11):2287–2296.
- Merianne R. Spencer, Matthew F. Garnett, and Ar-ialdi M. Miniño. 2024. Drug Overdose Deaths in the United States, 2002–2022. <https://www.cdc.gov/nchs/products/databriefs/db491.htm>. NCHS Data Brief, no. 491. National Center for Health Statistics. Accessed: 2026-02-25.
- Substance Abuse and Mental Health Services Administration. 2025. 2024 Companion Infographic Report: Results from the 2021 to 2024 National Surveys on Drug Use and Health. <https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health/national-releases>. SAMHSA Publication No. PEP25-07-006. Center for Behavioral Health Statistics and Quality. Accessed: 2026-02-25.
- Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction 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 15566–15589.
- Sarah E Wakeman, Marc R Larochelle, Omid Ameli, et al. 2020. Comparative effectiveness of different treatment pathways for opioid use disorder. *JAMA network open*, 3(2):e1920622.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Chenghao Yang, Tuhin Chakrabarty, Karli Hochstatter, et al. 2024. Identifying self-disclosures of use, misuse and addiction in community-based social media posts. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2507–2521.
- Zhou Yang, Spencer Bradshaw, Rattikorn Hewett, et al. 2022. Discovering opioid use patterns from social media for relapse prevention. *Computer*, 55(2):23–33.

A. Appendix

Model	Explanation	Dataset	Accuracy	F1
Original Results (Yang et al. 2024)				
DeBERTa-v3-Large	wo_explanation	Expert	0.676	0.657
		Novice	0.740	0.744
	w_explanation	Expert	0.738	0.726
		Novice	0.813	0.819
T5-3B	wo_explanation	Expert	0.635	0.612
		Novice	0.727	0.714
	w_explanation	Expert	0.644	0.655
		Novice	0.787	0.773
T5-11B	wo_explanation	Expert	0.714	0.704
		Novice	0.809	0.815
	w_explanation	Expert	0.766	0.770
		Novice	0.840	0.840
Reproduced Results (This Work)				
DeBERTa-v3-Base	wo_explanation	Expert	0.681	0.675
		Novice	0.707	0.712
	w_explanation	Expert	0.674	0.661
		Novice	0.733	0.739
DeBERTa-v3-Large	wo_explanation	Expert	0.708	0.707
		Novice	0.727	0.737
	w_explanation	Expert	0.735	0.739
		Novice	0.793	0.800
T5-3B	wo_explanation	Expert	0.690	0.703
		Novice	0.747	0.755
	w_explanation	Expert	0.738	0.750
		Novice	0.840	0.843
T5-11B	wo_explanation	Expert	0.732	0.739
		Novice	0.813	0.814
	w_explanation	Expert	0.772	0.788
		Novice	0.853	0.854

Table 8: Comparison of Original and Reproduced Baseline Results for DeBERTa and T5 Models

DeepSeek R1 Prompt

Classify the following post into one of these categories: "**Medical Use**", "**Misuse**", "**Addiction**", "**Recovery**", "**Relapse**", "**Not Using**"

- **Medical Use:** Medical use is defined as the use of prescription opioids that were prescribed by a medical professional for the purpose of treating a medical condition.
- **Misuse:** Misuse is defined as the use of a substance that does not follow medical indications or prescribed dosing. Substances are commonly used for nontherapeutic purposes to obtain psychotropic effects (euphoric, sedative, or anxiolytic). Misuse is not restricted to prescription opioids.
- **Addiction:** Addiction is defined as compulsive opioid use that occurs despite personal harm or negative consequences. It may involve impaired control and craving, neurobiologic dysfunction, physical and psychological dependence, and withdrawal.
- **Recovery:** Recovery is a process of change through which individuals improve their health and wellness, live a self-directed life, and strive to reach their full potential without using opioids.
- **Relapse:** Relapse is defined as the return to opioid use after an attempt to quit.
- **Not Using:** Posts should be labeled 'Not Using' when they discuss substances other than opioids, another person's opioid use, general opioid questions without evidence of personal use, or irrelevant information.

A reference answer is provided solely to help guide accurate reasoning toward the correct conclusion. However, the reasoning should not reference or rely on the presence of a reference answer in any explicit way.

Respond ONLY with a JSON object:

```
{  
  "reasoning": "Because ...",  
  "label": "Recovery"  
}
```

Post:
{ text }

Reference Answer:
{ label }

Figure 1: The prompt used with DeepSeek R1 to obtain reasoning traces along with the label

ICL-based relabeling prompt with GPT-5

Annotation Guidelines:

- **Medical Use:** Use of prescription opioids prescribed by a medical professional for a medical condition.
- **Misuse:** Use of a substance not following medical indications or prescribed dosing. Includes non-therapeutic use (e.g., euphoria, sedation). Misuse not restricted to prescription opioids.
- **Addiction:** Compulsive opioid use despite harm or negative consequences. Includes impaired control, craving, dependence, withdrawal.
- **Recovery:** Process of change where individuals stop using opioids, improve health and wellness, and live without opioids.
- **Relapse:** Return to opioid use after an attempt to quit.
- **Not Using:** Posts about substances other than opioids (e.g., marijuana), another person's opioid use, general questions without evidence of personal use, or irrelevant info.

Rules:

1. Output must be in JSON format: label: <one of [Medical Use, Misuse, Addiction, Recovery, Relapse, Not Using]>, rationale: <exact span(s) of text from the Reddit post that justify the label>
2. Rationale must be verbatim from the post (no paraphrasing).
3. Use the minimum sufficient span that clearly justifies the label.

Prompt Template:

You are a substance use research expert tasked with labeling Reddit posts about opioid use disorder (OUD).

{ guidelines }

Few-shot Examples

{ examples_str }

Task

Label the following post using the specified JSON output format and rules:

{ post_text }

Figure 2: The prompt used to relabel the posts using GPT-5 with few-shot ICL examples.

Simple zero-shot Prompt

Given the following title and text, please identify the appropriate opioid usage category among the following types: "**Medical Use**", "**Misuse**", "**Addiction**", "**Recovery**", "**Relapse**", "**Not Using**"

{ post_text }

Figure 3: The short prompt used with GPT-5 models in the zero-shot setting.

Guideline-enhanced zero-shot Prompt

You are a substance-use research expert tasked with labeling Reddit posts about opioid use disorder (OUD) into one of the following OUD stages.

- **Medical Use:** Medical use is defined as the use of prescription opioids that were prescribed by a medical professional for the purpose of treating a medical condition.
- **Misuse:** Misuse is defined as the use of a substance that does not follow medical indications or prescribed dosing. Substances are commonly used for nontherapeutic purposes to obtain psychotropic effects (euphoric, sedative, or anxiolytic). Misuse is not restricted to prescription opioids.
- **Addiction:** Addiction is defined as compulsive opioid use that occurs despite personal harm or negative consequences. It may involve impaired control and craving, neurobiologic dysfunction, physical and psychological dependence, and withdrawal.
- **Recovery:** Recovery is a process of change through which individuals improve their health and wellness, live a self-directed life, and strive to reach their full potential without using opioids.
- **Relapse:** Relapse is defined as the return to opioid use after an attempt to quit.
- **Not Using:** Posts should be labeled 'Not Using' when they discuss substances other than opioids, another person's opioid use, general opioid questions without evidence of personal use, or irrelevant information.

#Task

Determine the single OUD stage that describes the author's current situation at the time of writing. Choose exactly one: "**Medical Use**", "**Misuse**", "**Addiction**", "**Recovery**", "**Relapse**", "**Not Using**"

#Answer

Return only the label name on one line.

{ post_text }

Figure 4: The guideline-enhanced prompt used with GPT-5 models in zero-shot setting.

Representative opioid stage classification example where reasoning helps

Title: Opiates on amitriptyline

Post Text: Is there anybody who uses opiates on amitripty line? I used **morphine** like three times when on this drug (I take it for 2.5 months now, I take 60mg every evening) but even though I took **high dose** of **morphine** (on second try I had problems with breathing!) I don't feel **good buzz**, no satisfaction at all. Anyone knows how long do I have to give up the amitriptyline for the opiates to work? I have problems with sleeping and depression and amitriptyline helps me with it, but I miss my **morphine high** and at the moment it feels like waste of money.

Baseline prediction: Addiction

Reasoning-distilled prediction: Misuse. **Reasoning:** *The post discusses using morphine three times while on amitriptyline. The user is taking high doses of morphine (60mg) to achieve a 'good buzz', indicating non-therapeutic use for psychotropic effects. This aligns with the definition of Misuse, which involves using substances outside medical indications for purposes like achieving euphoria.*

Gold label: Misuse

Figure 5: Example illustrating reasoning leading to correct prediction compared with the baseline T5-3B.

Model	Stage	LR	Epochs	Batch	Grad Acc	Weight Decay
DeBERTa Base	Baseline	1e-5	15	8	1	–
	SCL Pretraining	3e-5	50	12	8	–
	SCL Finetuning	2e-5	10	8	1	–
DeBERTa Large	Baseline	1e-5	15	8	1	–
	SCL Pretraining	3e-5	50	12	8	–
	SCL Finetuning	2e-5	10	8	1	–
T5-3B	Baseline	1e-4	5	1	8	–
	SCL Pretraining	2e-5	50	12	8	–
	SCL Finetuning	1e-4	5	1	8	0.01
	Summarized Reasoning	1e-4	8	1	8	0.001
	Step-by-Step	1e-4	15	1	8	0.01
T5-11B	Baseline	5e-5	5	1	8	–
	SCL Pretraining	2e-5	50	12	8	–
	SCL Finetuning	1e-4	10	1	8	0.01
	Summarized Reasoning	8e-5	20	1	8	–
	Step-by-Step	5e-4	15	1	8	–

Table 9: Hyperparameter configurations for all models and training setups

FoodBench-QA: Overview of the Shared Task on Grounded Food and Nutrition Question Answering

Tome Eftimov^{1,2}, Ana Gjorgjevikj¹, Matej Martinc¹, Gjorgjina Cenikj¹,
Sašo Džeroski¹, Barbara Koroušić Seljak^{1,3}

¹ Jožef Stefan Institute, Jamova cesta 39, Ljubljana, Slovenia

² Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

³ Medical Faculty, University of Ljubljana, Vrazov trg 2, 1000 Ljubljana, Slovenia

{tome.eftimov, ana.gjorgjevikj, matej.martinc, gjorgjina.cenikj, saso.dzeroski, barbara.korousic}@ijs.si

Abstract

We present the results of the FoodBench-QA 2026 shared task at the CL4Health workshop, collocated with LREC 2026. FoodBench-QA challenges systems to answer food and nutrition questions using evidence from food composition databases and food-related ontologies. The shared task comprises three main tasks: nutrient estimation from recipe ingredients, evaluated using EU Regulation 1169/2011 tolerance thresholds; FSA traffic-light classification for fat, salt, saturates, and sugars; and food named entity recognition and linking to three ontologies, namely Hansard Taxonomy, FoodOn, and SNOMED CT. We received submissions from five participating teams across all tasks. For nutrient estimation, the best system achieved accuracy rates of 93.57% for protein, 86.50% for sugars, 84.65% for fat, and 86.26% for saturates. For FSA traffic-light prediction, the best macro F1 scores ranged from 0.65 to 0.90 across different nutrient-color combinations. For named entity linking, the best systems achieved macro F1 scores between 60.71% and 80.89% for natural text and 87.75% and 95.75% for artificial NEL datasets, depending on the ontology.

Keywords: question answering, nutrient estimation, food traffic light classification, named-entity linking

1. Introduction

Accurate food and nutrition information is essential for public health, dietary assessment, and personalized nutrition recommendations (Carlsen and Bruggemann, 2022; Agostoni et al., 2021; Guilpart et al., 2022). However, extracting and linking food-related information from text remains challenging due to the complexity of food terminology, the variety of ingredient descriptions, and the need to ground answers in authoritative food composition databases and food-related ontologies (Greenfield and Southgate, 2003; Gjorshoska et al., 2022).

The FoodBench-QA 2026 shared task addresses these challenges by providing a comprehensive evaluation framework for food and nutrition question answering systems. The task is designed to assess systems' abilities to estimate nutritional content (protein, sugars, fat, saturated fat) from recipe ingredients according to European Union (EU) regulatory standards (Bairati, 2017), classify food items according to the UK Food Standards Agency (FSA) traffic-light labeling system (Balcombe et al., 2010), and recognize and link food entities to standardized ontologies including Hansard Taxonomy (Abercrombie and Batista-Navarro, 2018), FoodOn (Doolley et al., 2018), and SNOMED CT (Donnelly et al., 2006).

The challenge builds upon the FoodBench dataset, a curated benchmark of question-answer pairs designed for training and evaluating large language models in food and nutrition domains. This shared task represents the first community evalu-

ation focused specifically on grounded food and nutrition question answering, following the tradition of successful BioNLP shared tasks.

FoodBench-QA 2026 was organized as part of the Third Workshop on Patient-Oriented Language Processing (CL4Health), collocated with LREC 2026: Language Resources and Evaluation Conference in Palma, Mallorca, Spain. The workshop provides a venue for computational linguistics research focused on patients' health and health-related issues concerning the public.

2. Background

To contextualize our approach, we summarize prior work in food and nutrition natural language processing (NLP), relevant ontologies, and applicable EU regulations.

Food and Nutrition NLP. Natural language processing for food and nutrition has gained increasing attention due to its applications in dietary assessment, food safety monitoring, and personalized nutrition (Cenikj et al., 2025). Previous work has addressed various aspects, including ingredient recognition (Stojanov et al., 2021, 2020; Cenikj et al., 2022a, 2020), recipe parsing (Popovski et al., 2019a,b; Ispirova et al., 2022), and nutrient estimation (Ispirova et al., 2024). However, standardized evaluation benchmarks have been lacking in this domain. Such benchmarks are essential in NLP and machine learning because they enable fair, reproducible, and transparent comparison of models

using shared datasets and metrics. They help track progress, identify strengths and weaknesses, and support scientific rigor through replicability. At the same time, they must be carefully designed and regularly updated to avoid overfitting and ensure real-world relevance.

Food Ontologies and Datasets. Three major semantic resources are used in this shared task. The **Hansard Taxonomy** is a hierarchical classification system for food products using structured codes such as AG.01.k [Flour], providing standardized categorization for food items (Parliament, 2022). **FoodOn** is a comprehensive ontology that covers food sources, food products, and their production processes, using URIs from the OBO (Open Biological and Biomedical Ontology) namespace and including links to NCBI Taxonomy for biological species (Dooley et al., 2018). **SNOMED CT**, the Systematized Nomenclature of Medicine Clinical Terms, includes food-related concepts relevant for clinical and medical applications, with URIs following the BioOntology format (Donnelly et al., 2006).

EU Regulation 1169/2011. The European Union Regulation 1169/2011 on food information to consumers establishes tolerance thresholds for nutritional declarations (Bairati, 2017). These thresholds define acceptable deviations between declared and actual nutrient values, providing a natural evaluation framework for nutrient estimation systems.

3. Task Descriptions

FoodBench-QA 2026 comprises three main tasks, with subtasks based on input modalities (Gjorgjevikj et al., 2026).

Task 1: Nutrient Estimation. The nutrient estimation task requires systems to predict the nutritional content per 100g for recipes based on their ingredients. Two subtasks were defined based on the input provided to systems (Neuhouser et al., 2023; Ispirova et al., 2024). In Task 1.1 (Ingredients Only), systems receive only the list of ingredients with quantities, while in Task 1.2 (Title + Ingredients), systems receive both the recipe title and the complete ingredient list. Four nutrients are evaluated: protein, sugars, fat, and saturated fat (saturates), with all values expressed in grams per 100g.

Task 2: FSA Traffic-Light Classification. The FSA traffic-light task requires systems to classify the nutritional quality of recipes according to the UK Food Standards Agency color-coded system (Balcombe et al., 2010; Kunz et al., 2020; Emrich et al.,

Nutrient	t_i	d_i
Protein, Sugars	≤ 10	≤ 2
	$10 < t_i \leq 40$	$\leq 0.2t_i$
	> 40	≤ 8
Fat	≤ 10	≤ 1.5
	$10 < t_i \leq 40$	$\leq 0.2t_i$
	> 40	≤ 8
Sat. Fat	< 4	≤ 0.8
	≥ 4	$\leq 0.2t_i$

Table 1: Tolerance rules used to evaluate nutrient estimation. t_i denotes the ground truth value, r_i the predicted value, and $d_i = |t_i - r_i|$ the absolute difference.

2017). Each of four nutrients (fat, salt, saturates, sugars) must be classified as green (low), orange (medium), or red (high). Similar to Task 1, two subtasks were defined: Task 2.1 performs classification based on the ingredient list and their quantities, while Task 2.2 uses the title, ingredients, and their quantities. This formulation results in a 12-class classification problem with four nutrients multiplied by three color categories.

Task 3: Named-Entity Recognition and Linking.

The NER+NEL task requires systems to identify food entities in text and link them to ontology concepts (Eftimov et al., 2017; Popovski et al., 2019a; Cenikj et al., 2020; Stojanov et al., 2021; Cenikj et al., 2022a; Gjorgjevikj et al., 2025). Two subtasks address different data sources. Task 3.1 (NER+NEL) is performed on two corpora: Scientific Abstracts (SA) containing technical food science literature (Cenikj et al., 2022b), and Food Composition Data (FCD) containing recipe descriptions (Ispirova et al., 2022). Task 3.2 (Artificial NEL) focuses on pure entity linking with pre-identified entities. In both subtasks, three ontologies have been considered: FoodOn, SNOMED CT, and Hansard Taxonomy.

4. Evaluation Metrics

Below we summarize the evaluation framework and metrics used to measure system effectiveness.

Task 1: Nutrient Estimation Metrics. Nutrient estimation is evaluated using tolerance thresholds defined in EU Regulation 1169/2011. A prediction is considered within tolerance based on specific rules for each nutrient type (see Table 1). The accuracy for each nutrient is calculated as the proportion of predictions falling within the tolerance threshold.

Task 2: FSA Traffic-Light Metrics. Each category-color combination is treated as a separate class, resulting in 12 classes total. Evaluation metrics include per-class Precision, Recall, and F1-Score.

Task 3: NER+NEL Metrics Entity linking is evaluated using two averaging methods. Macro Average calculates metrics per semantic tag and then averages them, giving equal weight to each semantic tag. Weighted Average weights metrics by the number of ground truth links per semantic tag. For each averaging method, we report Precision, Recall, and F1-Score. Entities with multiple tags are evaluated independently for each entity-link pair (This occurs due to the semantic nature of the tags (parent-child hierarchy) or the use of general food groups when the specific entity is not available).

5. Datasets

This section presents the datasets constructed for each task and summarizes their main statistics. Further details on the dataset construction process, the resources from which the datasets were derived, and the corresponding train-test splits are available in (Gjorgjevikj et al., 2026).

Task 1 and Task 2: Recipe Data. The nutrient estimation and FSA traffic-light datasets are derived from the Recipe1+ dataset (Marin et al., 2021). Table 2 reports the number of QA pairs in the datasets for Tasks 1 and 2.

Task	Training	Test
T1.1 – Ingredients Only	58,567	29,648
T1.2 – Title + Ingredients	58,652	29,649
T2.1 – Ingredients Only	58,585	29,578
T2.2 – Title + Ingredients	58,583	29,671

Table 2: Number of QA pairs in the FoodBench-QA T1 and T2 datasets.

Task 3.1: Natural Text NER+NEL. Two corpora were provided for the NER+NEL task on natural text. The Food Composition Data (FCD) contains 997 recipe descriptions with annotated food entities, split into 514 training recipes (51.6%) and 483 test recipes (48.4%) using a stratified split by recipe categories (Popovski et al., 2019b; Ispirova et al., 2022). The Scientific Abstracts (SA) corpus contains 470 scientific abstracts from food science literature, split into 329 training abstracts (70.0%) and 141 test abstracts (30.0%) using a random 70/30 split due to high category diversity (Cenikj et al., 2022b). Both datasets include annotations

for three ontologies: Hansard Taxonomy, FoodOn, and SNOMED CT.

Task 3.2: Artificial NEL. Three artificial entity linking datasets were created with pre-identified food entities. Artificial IR instances were generated to balance the distribution of food entities by identifying underrepresented entities for T3.1 and augmenting them until a minimum threshold of 150 mentions per entity was reached. The generated instances consist only of NEL instruction-response pairs linking sampled entity labels to FoodOn, Hansard, and SNOMED-CT, ensuring that all ontology entities are observed during fine-tuning while preventing duplicate examples to avoid data leakage (more details in (Gjorgjevikj et al., 2025)). The FoodOn NEL dataset is the largest with 13,492 entries, split into 9,445 training entries (70.0%) and 4,047 test entries (30.0%), containing 587 unique entities of which 42 are rare entities reserved for zero-shot evaluation. The SNOMEDCT NEL dataset contains 4,445 entries with a medical and clinical terminology focus, split into 3,112 training entries (70.0%) and 1,333 test entries (30.0%), with 310 unique entities including 30 rare entities. The Hansard NEL dataset contains 1,611 entries with the highest entity diversity, split into 915 training entries (56.8%) and 696 test entries (43.2%), featuring 1,082 unique entities of which 568 (52.6%) are rare entities, providing the most challenging zero-shot evaluation scenario.

All artificial NEL datasets employ stratified splits with rare entity protection, ensuring that entities appearing three or fewer times in the full dataset are exclusively placed in the test set for zero-shot evaluation. Table 3 summarizes the complete dataset statistics for the NER+NEL subtask.

6. Participating Systems

Five teams registered for FoodBench-QA 2026, with varying levels of participation across tasks. The best-ranked team (labeled as Team 1) participated in all tasks and submitted the most comprehensive set of runs, achieving the best results across most subtasks. The second-best team (labeled as Team 2) also participated in all tasks with competitive submissions. Team 3 focused on Task 1 (nutrient estimation), while Team 4 and Team 5 participated in the development phase. Participants were allowed to submit multiple runs to explore different configurations. The test phase ran from February 1-9, 2026, with submissions evaluated on the CodaBench platform.

⁰<https://www.codabench.org/competitions/12112/>

Dataset	Total	Train	Test	Unique Ent.	Rare Ent.
T3.1.1 – SA (Abstracts)	470	329	141	–	–
T3.1.2 – FCD (Recipes)	997	514	483	–	–
T3.2.1 – Artificial FoodOn	13,492	9,445	4,047	587	42
T3.2.2 – Artificial SNOMEDCT	4,445	3,112	1,333	310	30
T3.2.3 – Artificial Hansard	1,611	915	696	1,082	568
Total	21,015	14,315	6,700	–	–

Table 3: Summary of FoodBench-QA 2026 NER+NEL Datasets.

Applied methodologies. **Team 1:** For T1, a rule-based, evidence-grounded pipeline was developed in which heterogeneous ingredient measurements were resolved, quantities were standardized under SI standards (NIST SP 811), and interpretable, recipe-level nutrient values were deterministically estimated using publicly available resources (USDA FoodData Central and Recipe1M+). For T2 and T3, lightweight text-based classification was applied for FSA traffic-light prediction, and rule-based entity recognition was employed together with dictionary-driven ontology linking. **Team 2:** For T1, nutrient estimation and entity linking were framed as retrieval tasks rather than generative tasks, using TF-IDF and k-NN to identify the most similar recipes in the training database. To preserve authentic nutrient profiles, a conditional routing rule was implemented: when a retrieved recipe was nearly identical (over 95% similarity), ensembling was bypassed and the exact nutritional values were directly adopted. For standard matches below this threshold, a squared weighted voting mechanism was applied to ensure that high-confidence neighbors dominated the final prediction. For T2, an identical retrieval-augmented architecture to T1 was utilized. TF-IDF vectorization and weighted k-NN retrieval were applied to predict Food Standards Agency categories based on the closest neighbors. The same conditional routing rule was also applied to bypass ensembling and directly adopt categories from nearly identical matches (over 95% similarity). For T3.1 and 3.2 (NER and NEL), a unified Hybrid Regex-Dictionary approach was employed for both Named Entity Recognition and Linking. Recall was optimized using an automatic pluralization engine, while high precision was ensured through a Longest-Match-First regex strategy to prevent substring errors. The extracted entities were then mapped directly to the FoodOn, SNOMED CT, and Hansard ontologies using a high-precision dictionary, supported by an automated post-processing pipeline for strict formatting compliance. **Team 3:** For T1, multiple modeling strategies were evaluated to estimate recipe-level nutrients from unstructured ingredient lists, ranging from traditional lexical matching to large language models. A TF-IDF representation coupled with Ridge Regression was

used as a lightweight baseline, providing moderate estimation accuracy with near-instantaneous inference. To assess deeper semantic modeling, the DeBERTa-v3 encoder was evaluated, although its performance was limited by the scarcity of task-specific training data. In addition, few-shot inference with large language models (e.g., Gemma-3-27B) and a hybrid refinement pipeline combining TF-IDF retrieval with Gemini 2.5 Flash were explored. These approaches leverage pre-trained world knowledge to implicitly perform ingredient disambiguation and unit normalization, achieving higher accuracy under the strict tolerance criteria of EU Regulation 1169/2011, albeit at the cost of increased computational latency. For T2 and T3, lightweight text-based classification was applied for FSA traffic-light prediction, while rule-based entity recognition combined with ontology-based entity linking was used to map extracted food entities to controlled vocabularies. **Baseline:** To assess whether LLM-based solutions show promising signals when trained on the data, a previous study reports results for each task (Gjorgjevikj et al., 2026). However, these results should be interpreted cautiously, as the evaluations were not performed using identical training and test splits.

7. Results

We present the task-wise results and compare the performance of the submitted systems.

Task 1: Nutrient Estimation Results. Table 4 presents the results for the nutrient estimation task, where accuracy represents the percentage of predictions within EU Regulation 1169/2011 tolerance thresholds.

The best results were achieved by “Team 1”, with accuracy rates exceeding 84% for all nutrients when using title and ingredients (T1.2) (see Figure 1). Adding recipe titles provided marginal improvements over ingredients-only predictions, with the largest gain observed for sugars at 0.06 percentage points.

Task	Team	Protein	Sugars	Fat	Saturates
T1.1	Team 1 (best) (six entries)	93.44	86.44	84.46	86.18
	Team 2 (a single entry)	76.35	69.66	67.45	71.55
	Team 3 (two entries)	59.71	43.03	39.21	47.13
T1.2	Team 1 (best) (six entries)	93.57	86.50	84.65	86.26
	Team 2 (a single entry)	77.73	71.67	69.28	73.41

Table 4: Task 1 Results: Nutrient Estimation Accuracy (%)

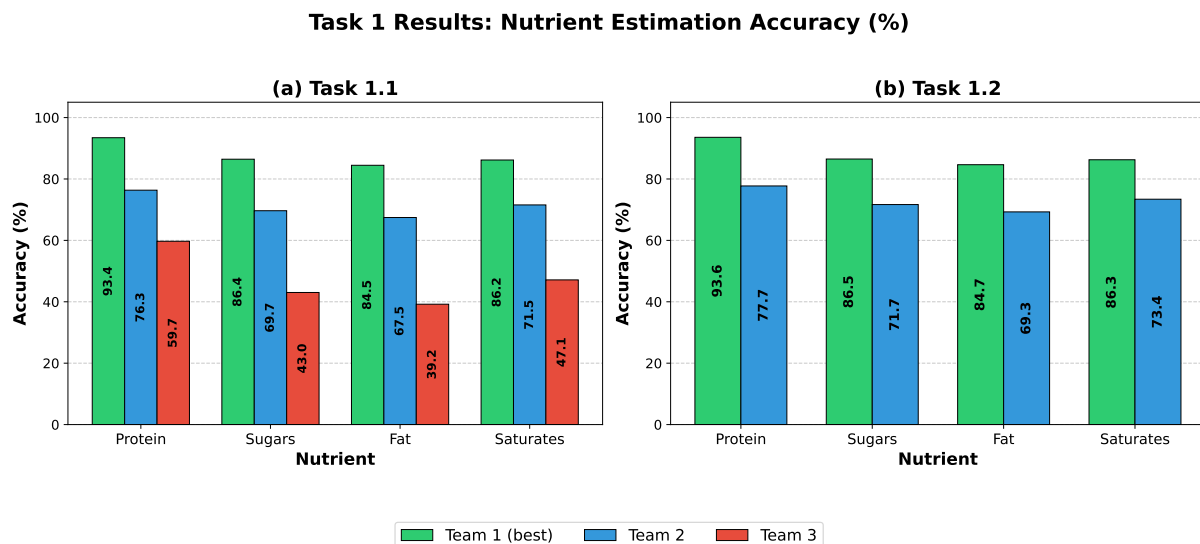


Figure 1: Task 1 results: Nutrient estimation accuracy (%) across participating teams.

Task 2: FSA Traffic-Light Results. Table 5 presents the F1-scores for each nutrient-color combination, showing results for the best-performing submissions (also see Figure 2). The results show that green (low) classifications are generally easier to predict than orange (medium) and red (high) classifications. The orange category consistently shows the lowest F1-scores across all nutrients, suggesting that medium-level classifications are most challenging. Adding recipe titles improved classification performance across most categories. Across both teams, the best overall performance was achieved by Team 2 in Task 2.1 (notably for Saturates and Salt), while Team 1 dominated Task 2.2, consistently obtaining the highest F1-scores across most nutrients and traffic-light classes when title information was included.

Task 3.1: NER+NEL on Natural Text Results Table 6 presents the entity linking results for natural text datasets.

The results demonstrate that SNOMED CT linking achieves the highest F1-scores for both SA and FCD datasets, followed by FoodOn and Hansard. The higher precision on FCD compared to SA suggests that recipe text is easier to process than scientific abstracts. Across both subtasks, clear per-

formance differences emerge between the two teams. On T3.1.1 (Scientific Abstracts), Team 1 consistently outperforms Team 2 in terms of Weighted F1 across all three ontologies, with particularly strong results for SNOMED CT (F1 = 80.89), where it achieves both the highest precision and recall. Team 2 remains competitive in precision, especially for FoodOn and SNOMED CT, but lower recall limits its overall F1 performance on this dataset. In contrast, on T3.1.2 (Food Composition Data), Team 2 demonstrates stronger overall performance, achieving the highest Weighted F1 scores for all three ontologies, including the best overall result (F1 = 80.89 on SNOMED CT). While Team 1 maintains strong recall, particularly for Hansard and SNOMED CT, Team 2’s higher precision across ontologies leads to superior overall F1 performance. Notably, SNOMED CT emerges as the most robust ontology across both datasets and teams, consistently yielding the highest scores (see Figure 3).

Task 3.2: Artificial NEL Results. Table 7 presents the entity linking results for the artificial NEL datasets.

The artificial NEL task shows significantly higher performance than natural text NER+NEL (see Figure 4). In Task 3.2 (Artificial NEL), Team 1 clearly

Task 2 Results: FSA Traffic-Light F1-Scores

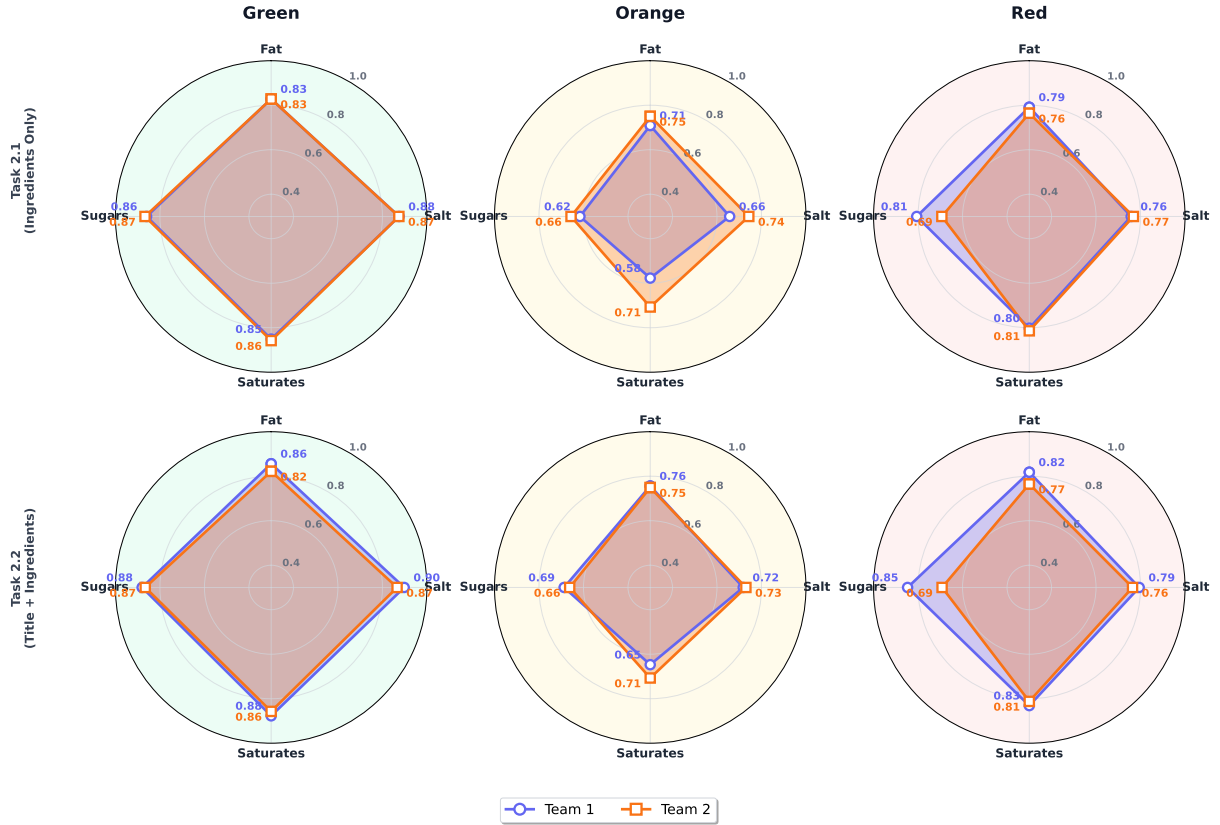


Figure 2: Task 2 results: FSA traffic-light classification F1-scores across nutrients and teams.

Nutrient	Team	Green	Orange	Red
<i>Task 2.1 (Ingredients Only)</i>				
Fat	Team 1	0.828	0.709	0.791
Salt	Team 1	0.875	0.657	0.758
Saturates	Team 1	0.851	0.577	0.802
Sugars	Team 1	0.860	0.616	0.806
Fat	Team 2	0.828	0.750	0.764
Salt	Team 2	0.874	0.743	0.768
Saturates	Team 2	0.858	0.707	0.814
Sugars	Team 2	0.867	0.656	0.693
<i>Task 2.2 (Title + Ingredients)</i>				
Fat	Team 1	0.856	0.756	0.818
Salt	Team 1	0.898	0.718	0.794
Saturates	Team 1	0.877	0.647	0.831
Sugars	Team 1	0.880	0.687	0.847
Fat	Team 2	0.822	0.749	0.765
Salt	Team 2	0.866	0.730	0.764
Saturates	Team 2	0.857	0.707	0.812
Sugars	Team 2	0.866	0.662	0.692

Table 5: Task 2 Results: FSA Traffic-Light F1-Scores

outperforms Team 2 across all ontologies and evaluation metrics. Team 1 achieves substantially higher recall, particularly for FoodOn (94.80) and SNOMED CT (98.90), which translates into consistently superior F1 scores. The largest performance gap is observed for Hansard, where Team 1 reaches an F1 of 87.75 compared to 56.20 for Team 2. Although Team 2 remains competitive in precision—especially for FoodOn and SNOMED CT—the significantly lower recall limits its overall effectiveness.

8. Analysis of Results

The following analysis highlights major trends and insights derived from the experimental results.

Task Difficulty Analysis: The results reveal a clear hierarchy of task difficulty. Artificial NEL proved to be the easiest task, with weighted F1-scores reaching up to 95.75%, benefiting from pre-identified entities that eliminate the need for entity recognition. Nutrient estimation achieved accuracy rates up to 93.57%, leveraging well-defined tolerance thresholds. FSA traffic-light classification showed F1-scores ranging from 0.61 to 0.90,

Dataset	Ontology	Team	Precision	Recall	F1
<i>T3.1.1 – Scientific Abstracts (SA)</i>					
SA	Hansard	Team 1	54.11	71.51	60.71
SA	Hansard	Team 2	47.99	38.38	41.13
SA	FoodOn	Team 1	74.53	71.48	71.38
SA	FoodOn	Team 2	75.16	54.23	61.43
SA	SNOMEDCT	Team 1	83.90	84.34	80.89
SA	SNOMEDCT	Team 2	83.90	69.40	74.32
<i>T3.1.2 – Food Composition Data (FCD)</i>					
FCD	Hansard	Team 1	74.35	77.06	75.09
FCD	Hansard	Team 2	81.04	75.60	77.75
FCD	FoodOn	Team 1	91.15	70.39	77.92
FCD	FoodOn	Team 2	94.67	69.98	79.01
FCD	SNOMEDCT	Team 1	91.30	72.65	79.42
FCD	SNOMEDCT	Team 2	95.10	72.51	80.89

Table 6: Task 3.1 Weighted Results: NER+NEL on Natural Text (Best Submissions)

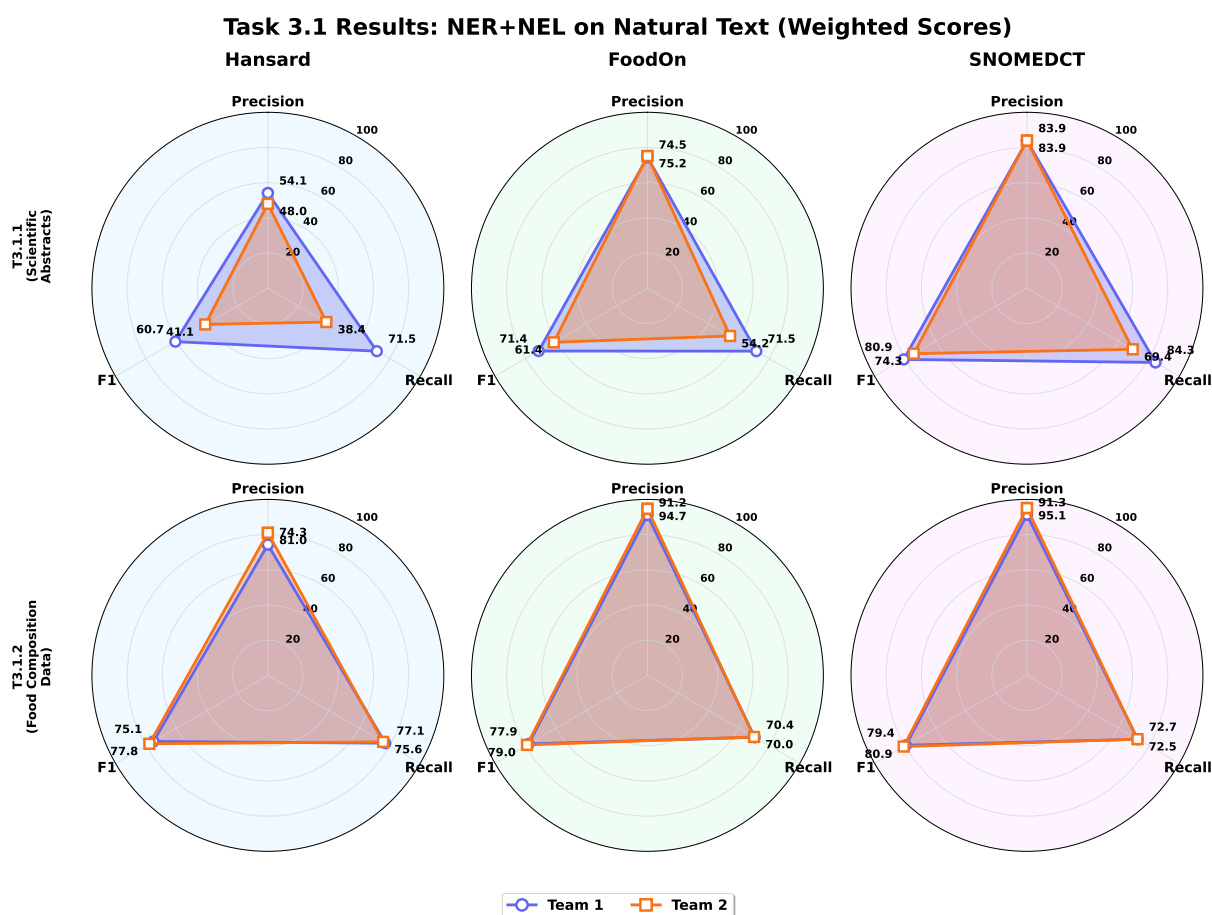


Figure 3: Performance evaluation of named entity recognition and linking on natural text: A comparison across Scientific Abstracts (T3.1.1) and Food Composition Data (T3.1.2).

with orange classifications proving most challenging across all nutrients. Natural text NER+NEL emerged as the hardest task with F1-scores up to 80.89%, as it requires both accurate entity recognition and correct linking to ontology concepts.

Impact of Recipe Titles: Comparing the ingredients-only subtasks (T1.1 and T2.1) with the title plus ingredients subtasks (T1.2 and T2.2) reveals the impact of additional contextual information. Nutrient estimation showed marginal

Dataset	Ontology	Team	Precision	Recall	F1
<i>Task 3 – Natural Text Evaluation</i>					
Artificial text	Hansard	Team 1	90.53	85.89	87.75
Artificial text	Hansard	Team 2	75.48	45.86	56.20
Artificial text	FoodOn	Team 1	86.35	94.80	90.04
Artificial text	FoodOn	Team 2	92.49	64.56	74.71
Artificial text	SNOMED CT	Team 1	93.07	98.90	95.75
Artificial text	SNOMED CT	Team 2	93.22	72.01	80.68

Table 7: Task 3.2 Weighted Results: Artificial NEL (Best Submissions).

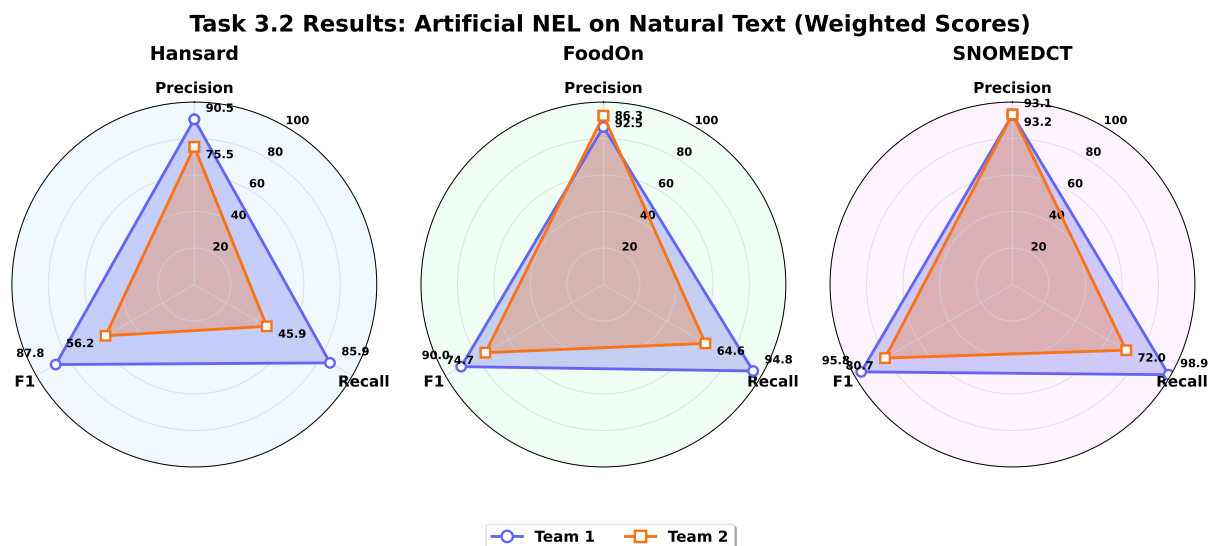


Figure 4: Performance analysis of named entity linking on artificially generated text using Hansard, FoodOn, and SNOMED CT ontologies.

improvements of approximately 0.13 percentage points for protein and 0.06 percentage points for sugars. FSA traffic-light classification demonstrated more substantial improvements, particularly for red classifications. Recipe titles provide contextual information that helps disambiguate ingredient interpretations and improve overall prediction accuracy.

Ontology-Specific Performance: Across both natural text (scientific abstracts and recipe description) and artificial NEL tasks, consistent patterns emerged in ontology-specific performance. SNOMED CT consistently achieved the highest F1-scores, likely due to more standardized medical terminology that facilitates accurate entity linking. FoodOn showed strong performance with good coverage of food items across diverse categories.

Zero-Shot Evaluation: The artificial NEL datasets were specifically designed with rare entities (appearing three or fewer times) exclusively in the test set to enable zero-shot evaluation. The FoodOn dataset contains 42 rare entities,

SNOMED CT contains 30 rare entities, and Hansard contains 568 rare entities representing 52.6% of its unique entities. The Hansard dataset provides the most challenging zero-shot scenario due to its high proportion of rare entities, testing systems' ability to generalize to previously unseen food terminology.

9. Conclusions and Future Work

FoodBench-QA 2026 established the first community benchmark for grounded food and nutrition question answering, attracting five teams and providing insights into the capabilities and limitations of current food NLP systems. It introduced a multi-task evaluation framework for nutrient estimation, FSA classification, and entity linking, with datasets, evaluation scripts, and submissions publicly available on CodaBench <https://www.codabench.org/competitions/12112/> and Zenodo at <https://doi.org/10.5281/zenodo.17798877>.

10. Acknowledgments

The authors acknowledge the support of the Slovenian Research Agency through program grants No. P2-0098, No. and P2-0103, project grants No. J7-70265, No. GC-0001 and No. PR-12393. This work is also funded by the European Union under Grant Agreement 101211695 (HE MSCA-PF AutoLLMSelect), Grant Agreement 101187010 (HE ERA Chair AutoLearn SI), Grant Agreement 101198470 (LLMs4EU), Grant Agreement 101060712 (HE FishEUTrust), and Grant Agreement 101214398 (ELLIOT). We also acknowledge the support of the EC/EuroHPC JU and the Slovenian Ministry of HESI via the project SLAIF (grant number 101254461). We thank all participating teams for their contributions to FoodBench-QA 2026. We also thank the CL4Health workshop organizers for hosting the shared task. We acknowledge the participating teams based on the Codabench platform: Team 1 (rbls-lab), Team 2 (raviranj), and Team 3 (andybox111).

11. Ethics Statement and Broader Impact

Intended Use and Societal Impact. The FoodBench-QA shared task aims to advance research in grounded food and nutrition question answering, with potential applications in dietary assessment, public health monitoring, and personalized nutrition support. By promoting standardized evaluation benchmarks and transparent comparison of methods, the task contributes to the development of reliable AI systems for food and health-related information processing. However, the presented systems are intended for research purposes and should not be used as substitutes for professional medical or nutritional advice.

Risks and Misuse. Automated nutrient estimation and food classification systems may produce inaccurate predictions, particularly in cases of incomplete ingredient information, ambiguous food descriptions, or unseen entities. Incorrect nutritional estimates or classifications could potentially mislead users if deployed without appropriate validation or human oversight. We therefore emphasize that outputs should be interpreted with caution and require expert verification in real-world health or clinical applications.

Data Sources and Privacy. The datasets used in the shared task are derived from publicly available resources, including recipe datasets, scientific abstracts, and food ontologies. These datasets do not contain personally identifiable information,

and no personal or sensitive user data were collected or processed. The annotation and dataset construction processes followed standard research practices to ensure responsible data handling.

Bias and Representation. The datasets primarily reflect food terminology, recipes, and nutritional standards derived from specific cultural and regulatory contexts, including European Union nutrition regulations and English-language resources. As a result, the benchmark may not fully represent global dietary practices, regional cuisines, or multilingual food descriptions. Systems trained or evaluated on this benchmark may therefore exhibit reduced performance for underrepresented food concepts or cultural contexts.

Reproducibility and Transparency. To promote transparency and responsible research, all datasets, evaluation scripts, and submission results are publicly released through CodaBench and Zenodo. This enables independent validation, comparison of methods, and further research on trustworthy food and nutrition AI systems.

12. Limitations

The FoodBench-QA shared task has several limitations that should be considered when interpreting the results.

First, participation was limited to a small number of teams, and not all teams submitted results for every task or subtask. This restricts the diversity of methodological approaches and limits the generalizability of performance comparisons.

Second, the benchmark focuses on structured recipe data and curated text corpora, which may not fully capture the complexity and variability of real-world food descriptions, informal language, or multilingual settings.

Third, nutrient estimation is evaluated using tolerance thresholds defined by EU Regulation 1169/2011, which provide a regulatory evaluation framework but may not reflect all practical requirements for dietary assessment or clinical decision-making.

Fourth, the artificial named entity linking datasets simplify the problem by providing pre-identified entities, which do not fully reflect the challenges of end-to-end information extraction from natural text.

Finally, although the benchmark covers multiple ontologies, it remains limited to three semantic resources and does not account for the full diversity of food-related knowledge representations.

Future work should address these limitations by expanding dataset diversity, increasing participation, incorporating multilingual and culturally di-

verse food data, and evaluating systems in real-world deployment scenarios.

13. Bibliographical References

- Gavin Abercrombie and Riza Theresa Batista-Navarro. 2018. A sentiment-labelled corpus of hansard parliamentary debate speeches. In *International Language Resource and Evaluation Conference 2018: ParlaCLARIN Workshop*, pages 43–47. Clarin.
- C Agostoni, Stefania Boccia, S Banni, PM Mannucci, and A Astrup. 2021. Sustainable and personalized nutrition: From earth health to public health. *European Journal of Internal Medicine*, 86:12–16.
- Lorenzo Bairati. 2017. The food consumer's right to information on product country of origin: trends and outlook, beyond eu regulation 1169/2011. *Journal of European Consumer and Market Law*, 6(1).
- Kelvin Balcombe, Iain Fraser, and Salvatore Di Falco. 2010. Traffic lights and food choice: A choice experiment examining the relationship between nutritional food labels and price. *Food policy*, 35(3):211–220.
- Lars Carlsen and Rainer Bruggemann. 2022. The 17 united nations' sustainable development goals: A status by 2020. *International Journal of Sustainable Development & World Ecology*, 29(3):219–229.
- Gjorgjina Cenikj, Mauro Dragoni, Tome Eftimov, Barbara KOROU ŠI C SELJAK, Agnieszka Ławrynowicz, Fnu Mohbat, Oshani Seneviratne, Yoko Yamakata, and Mohammed J Zaki. 2025. Neurosymbolic methods for food computing. In *Handbook on Neurosymbolic AI and Knowledge Graphs*, pages 1019–1056. IOS Press.
- Gjorgjina Cenikj, Gašper Petelin, Barbara Koroušić Seljak, and Tome Eftimov. 2022a. Scifoodner: Food named entity recognition for scientific text. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4065–4073. IEEE.
- Gjorgjina Cenikj, Gorjan Popovski, Riste Stojanov, Barbara Koroušić Seljak, and Tome Eftimov. 2020. Butter: Bidirectional lstm for food named-entity recognition. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3550–3556. IEEE.
- Gjorgjina Cenikj, Eva Valenčič, Gordana Ispirova, Matevž Ogrinc, Riste Stojanov, Peter Korošec, Ermanno Cavalli, Barbara Koroušić Seljak, and Tome Eftimov. 2022b. Cafeterias corpus: scientific abstracts annotated across different food semantic resources. *Database*, 2022:baac107.
- Kevin Donnelly et al. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.
- Damion M Dooley, Emma J Griffiths, Gurinder S Gosal, Pier L Buttigieg, Robert Hoehndorf, Matthew C Lange, Lynn M Schriml, Fiona SL Brinkman, and William WL Hsiao. 2018. Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*, 2(1):23.
- Tome Eftimov, Barbara Koroušić Seljak, and Peter Korošec. 2017. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PloS one*, 12(6):e0179488.
- Teri E Emrich, Ying Qi, Wendy Y Lou, and Mary R L'Abbe. 2017. Traffic-light labels could reduce population intakes of calories, total fat, saturated fat, and sodium. *PloS one*, 12(2):e0171188.
- Ana Gjorgjevikj, Matej Martinc, Gjorgjina Cenikj, Sašo Džeroski, Barbara Koroušić Seljak, and Tome Eftimov. 2025. Foodsem: Large language model specialized in food named-entity linking. In *International Conference on Discovery Science*, pages 395–410. Springer.
- Ana Gjorgjevikj, Matej Martinc, Gjorgjina Cenikj, Riste Stojanov, Jan Drole, Gordana Ispirova, Giulia Menichetti, Nives Ogrinc, Dimitar Trajanov, Sašo Džeroski, Barbara Koroušić Seljak, and Tome Eftimov. 2026. Large language models in food and nutrition science: Opportunities, challenges, and the case of foodylm. *Current Research in Food Science*. In press (accepted February 13, 2026).
- Ivana Gjørshoska, Tome Eftimov, and Dimitar Trajanov. 2022. Missing value imputation in food composition data with denoising autoencoders. *Journal of Food Composition and Analysis*, 112:104638.
- Heather Greenfield and David AT Southgate. 2003. *Food composition data: production, management, and use*. Food & Agriculture Org.
- Nicolas Guilpart, Toshichika Iizumi, and David Makowski. 2022. Data-driven projections suggest large opportunities to improve europe's soybean self-sufficiency under climate change. *Nature Food*, 3(4):255–265.

- Gordana Ispirova, Gjorgjina Cenikj, Matevž Ogrinc, Eva Valenčič, Riste Stojanov, Peter Korošec, Ermanno Cavalli, Barbara Koroušić Seljak, and Tome Eftimov. 2022. Cafeteriafcd corpus: food consumption data annotated with regard to different food semantic resources. *Foods*, 11(17):2684.
- Gordana Ispirova, Tome Eftimov, Sašo Džeroski, and Barbara Koroušić Seljak. 2024. Msgen: Measuring generalization of nutrient value prediction across different recipe datasets. *Expert Systems with Applications*, 237:121507.
- Sonja Kunz, Simona Haasova, Jannik Rieß, and Arnd Florack. 2020. Beyond healthiness: the impact of traffic light labels on taste expectations and purchase intentions. *Foods*, 9(2):134.
- Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2021. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):187–203.
- Marian L Neuhouser, Ross L Prentice, Lesley F Tinker, and Johanna W Lampe. 2023. Enhancing capacity for food and nutrient intake assessment in population sciences research. *Annual review of public health*, 44(1):37–54.
- THIRTEENTH Parliament. 2022. Hansard. URL <https://hansard.parliament.uk>.
- Gorjan Popovski, Stefan Kochev, Barbara Korousic-Seljak, and Tome Eftimov. 2019a. Foodie: A rule-based named-entity recognition method for food information extraction. *ICPRAM*, 12:915.
- Gorjan Popovski, Barbara Koroušić Seljak, and Tome Eftimov. 2019b. Foodbase corpus: a new resource of annotated food entities. *Database*, 2019:baz121.
- Riste Stojanov, Gorjan Popovski, Gjorgjina Cenikj, Barbara Koroušić Seljak, and Tome Eftimov. 2021. A fine-tuned bidirectional encoder representations from transformers model for food named-entity recognition: Algorithm development and validation. *Journal of Medical Internet Research*, 23(8):e28229.
- Riste Stojanov, Gorjan Popovski, Nasi Jofce, Dimitar Trajanov, Barbara Koroušić Seljak, and Tome Eftimov. 2020. Foodviz: Visualization of food entities linked across different standards. In *Machine Learning, Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, July 19–23, 2020, Revised Selected Papers, Part II* 6, pages 28–38. Springer.

Overview of the CT-DEB'26 Shared Task on Predicting Dosing Errors in Interventional Clinical Trials

Sohrab Ferdowsi¹, Félicien Hêche¹, Anthony Yazdani¹,
Edward Choi², Sara Sansaloni-Pastor³, Douglas Teodoro¹

¹ Department of Radiology and Medical Informatics, University of Geneva, Switzerland

² KAIST, Republic of Korea

³ Actelion Pharmaceuticals Ltd, Basel, Switzerland

Abstract

Dosing errors represent an important source of medication-related risk in interventional clinical trials, potentially affecting both participant safety and the validity of study outcomes. Despite their importance, systematic methods for predicting dosing error risk from trial design information remain largely unexplored. To address this gap, we organized the *Clinical Trial Dosing Error Benchmark 2026* (CT-DEB'26) shared task, hosted at the CL4Health workshop at LREC 2026. The task focuses on predicting the risk of dosing errors in interventional clinical trials using heterogeneous information extracted from ClinicalTrials.gov, including structured protocol metadata and long-form textual descriptions. The released benchmark dataset contains over 42,000 clinical trial records spanning multiple study phases and therapeutic areas, annotated with binary labels indicating a significant high rate of dosing errors. Participants were asked to develop ML models capable of estimating trial-level dosing error risk, evaluated primarily using the ROC-AUC metric under strong class imbalance. The shared task was conducted in two phases and attracted 15 submissions in the development stage and 4 submissions in the final evaluation phase. This paper provides an overview of the shared task, describing the dataset construction, evaluation protocol, and participating systems. In addition, we present a schema-aware CatBoost baseline that leverages structured trial metadata and simple textual statistics, achieving ROC-AUC scores of 0.8606 and 0.8624 on the Phase 1 and Phase 2 leaderboards, respectively. We further summarize the approaches proposed by participating teams, which explore both feature-engineering pipelines and transformer-based text representations. The results highlight the importance of structured trial design variables and hybrid modeling strategies combining tabular and textual information. Finally, we discuss limitations of the benchmark and outline future directions for applying natural language processing and ML to improve medication safety in clinical trial design.

Keywords: clinical trials, medication errors, dosing errors, shared task, clinical NLP, machine learning

1. Introduction

1.1. Clinical Trials

Interventional clinical trials (CTs) constitute the cornerstone of pharmaceutical research and development, providing the evidentiary basis for assessing the safety and efficacy of medicinal products and interventions. Once the drug discovery research and pre-clinical animal-based studies are successful, for a candidate medication to reach the market, regulatory agencies require various phases of clinical studies on human volunteers, each designed to rigorously evaluate specific aspects of the drug's performance and safety profile.

Various regulations such as the (U.S. Food and Drug Administration, 2024), the (European Parliament and Council of the European Union, 2014), as well as the (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), 2016) require careful planning of the CTs prior to their execution in what is known as CT protocols (Cipriani and Barbui, 2010). Despite these efforts and the oversight by regulatory agencies, CTs remain costly, lengthy, and specifically highly failure-prone processes. Various studies es-

timate that no more than 14% of candidate drugs successfully progress from phase I to market authorization (Wong et al., 2018; Sun et al., 2022), while CT execution constitutes a major portion of the very costly (approximately 1.3B\$ on the average) drug R&D life cycle (Wouters et al., 2020; Mulcahy et al., 2025). These failures are often attributed not only to the lack of drug efficacy, but also to operational and procedural deficiencies in trial execution (Martin et al., 2017). As a result, identifying and mitigating such inefficiencies early in the trial lifecycle is of critical importance for cost reduction of clinical studies and hence the drug prices and eventually the sustainability of the healthcare systems, as well as for the safety of participants of these studies.

Towards optimizing CT design, large public repositories such as ClinicalTrials.gov (CTGov) (National Library of Medicine (US), 2000) provide access to the past trial protocols and outcomes, enabling data-driven methods to estimate trial-level risks and to identify protocol design factors associated with operational failure. This is furthermore motivated by the advances and success stories of methods based on Machine Learning (ML) and Artificial Intelligence (AI) in healthcare. To get a

better overview of the application of these methods to trial optimization, the reader is referred to the recent scoping review by (Teodoro et al., 2025), where various categories of risk prediction models are reviewed.

1.2. Dosing Errors

On the other hand, medication errors, and in particular dosing errors, are among the major sources of adverse events imposing substantial burden on healthcare systems (Elliott et al., 2021; Rouhani et al., 2018; Hodkinson et al., 2020; Tariq et al., 2018). Within the context of clinical studies, medication errors can compromise trial validity by biasing efficacy estimates (Vrijens et al., 2024), jeopardize participant safety (Tariq et al., 2025), and lead to regulatory non-compliance or trial termination (European Medicines Agency, 2023).

Medication errors are defined as failures in the treatment process that lead to, or have the potential to lead to, harm to the patient (Aronson, 2009). Examples of dosing-related medication errors include incorrect dose amount, frequency, timing, or administration route.

While medication safety has been extensively studied in post-marketing and routine clinical care settings, comparatively little attention has been paid to the systematic analysis and prediction of dosing errors occurring during clinical research itself.

Prior work has investigated trial outcome prediction and failure risk estimation using protocol-level information (Ferdowsi et al., 2023, 2021), as well as the prediction of adverse drug events from clinical trial data (Yazdani et al., 2025), highlighting the value of data-driven approaches for assessing trial-level risk. However, as emphasized by the recent scoping review of (Hêche et al., 2026), ML-based studies of medication errors have almost exclusively focused on errors occurring in routine clinical practice, leaving a notable gap with respect to errors arising in the conduct of CTs. Addressing this gap is particularly challenging due to the heterogeneity of trial designs, the rarity of documented dosing errors, and the prevalence of unstructured textual information in trial protocols and reports.

1.3. Shared Task

To foster community-driven progress on this problem, we organized the *Clinical Trial Dosing Error Benchmark 2026* (CT-DEB'26), an open shared task aimed at predicting the risk of dosing errors in interventional clinical trials using information available prior to trial initiation. The shared task leverages a large, publicly released dataset derived from the ClinicalTrials.gov registry, combining structured trial design variables with long-form textual

descriptions extracted from registry entries and protocol documents. Participants were invited to develop machine learning models that, given only pre-initiation protocol information, estimate the probability that a trial will exhibit an elevated rate of dosing errors during its execution. System performance was evaluated primarily using the area under the receiver operating characteristic curve (ROC-AUC), among other complementary metrics.

CT-DEB'26 was hosted on the CodaBench platform (Xu et al., 2022) and conducted in two phases, supporting model development and subsequent evaluation on a fully held-out test set. The challenge was organized as part of the CL4Health workshop at LREC 2026 and aims to establish a reusable benchmark for studying dosing error prediction in clinical trials.

This paper provides an overview of the CT-DEB'26 shared task. We first describe the dataset construction process and the prediction task formulation in Section 2. We then present the challenge setup in Section 3, including the task definition and evaluation protocol. Section 4 summarizes the submitted systems and presents a baseline model. Finally, Section 5 discusses the main findings and directions for future research.

2. Dataset

2.1. Data Source and Trial Selection

The CT-DEB'26 shared task is based on the `ct-dosing-errors-benchmark` dataset (Hêche et al., 2026) derived from completed and terminated interventional clinical trials registered at the CTGov registry (National Library of Medicine (US), 2000), which was the main source of the raw data used due to its scale, standardized reporting, as well as its regulatory relevance. The dataset was curated to support trial-level prediction of dosing error risk using only information available prior to the trial initiation, thereby reflecting realistic deployment conditions for prospective risk assessment.

Trials were selected according to predefined inclusion criteria, including interventional study design, availability of reported results, and a documented completion date. Restricting the dataset to completed or terminated trials ensures that dosing error outcomes can be derived from finalized adverse event reports, while enabling temporally coherent dataset splitting. A total of 42,112 trials satisfied these criteria and were included in the full dataset.

The dataset is publicly released on the Hugging Face Hub¹. A detailed account of the dataset

¹The shared-task dataset is available at <https://huggingface.co/datasets/sssohrab/ct-dosing-errors-benchmark>. A

Feature	Avail.	Mean	Min	Max
enrollmentCount	100%	263.81	1	864,493
numArms	100%	2.33	0	43
numInterventions	100%	2.48	1	44
numLocations	100%	22.88	0	1,611

Table 1: Summary statistics of numerical features.

construction pipeline, feature engineering, labeling methodology and splitting strategy is provided in the companion paper (Hêche et al., 2026). Below we provide a summary as is relevant for the shared task.

2.2. Features

For each trial, a heterogeneous set of features was extracted, comprising numerical, categorical, and textual variables. Structured features describe core aspects of trial design and logistics, such as trial phase, enrollment size, intervention configuration, allocation and masking strategies, sponsor and oversight information, and characteristics of treatment arms and interventions.

In addition to structured variables, the dataset includes multiple free-text fields drawn from registry entries and protocol-related documents, such as brief summaries, detailed descriptions, arm and intervention descriptions, and, when available, text extracted from protocol PDF files. Approximately 42% of trials contain long-form textual descriptions, which often encode nuanced procedural details relevant to medication administration and dosing. As a result, effective utilization of unstructured text constitutes a central challenge of the benchmark.

To ensure consistency and reproducibility, all features were validated and parsed using structured data schemas with tools such as the Pydantic library (Pydantic AI), or Python’s standard data types, such as the enumerated data structures (Enum) (Python Software Foundation). Notably, this results in categorical variables constraint to predefined value sets and would later facilitate ML-related feature engineering. These schemas were created according to the structures defined in CT-Gov’s instructions.

Tables 1, 2, 3, and 4 provide summary statistics for the numerical, textual, categorical, and list-valued categorical features, respectively.

Notably, among the textual fields, `protocolPdfText` is both the sparsest and by far the longest, reflecting the partial availability of full protocol documents and the substantial additional pro-

more detailed version with additional risk categories is available at <https://huggingface.co/datasets/ds4dh/ct-dosing-errors>. The source code used to reproduce these datasets is available at <https://github.com/ds4dh/CT-dosing-errors>.

cedural detail they contain when present.

Overall, the dataset combines structured numerical and categorical protocol descriptors with multiple free-text fields of highly variable length, yielding a multimodal benchmark for trial-level dosing error prediction.

2.3. Outcome Labeling and Dataset Splits

The prediction target is a binary variable indicating whether a trial exhibited an unusually high rate of dosing errors during its execution. Labels were derived from adverse event (AE) reports available in the ClinicalTrials.gov registry, which are coded using the Medical Dictionary for Regulatory Activities (MedDRA, version 27.1) (MedDRA Maintenance and Support Services Organization, 2025).

To identify dosing-related events, MedDRA categories associated with overdosing, underdosing, and medication-use errors were used as entry points in the MedDRA hierarchy. The descendant terms under these categories were reviewed by clinical pharmacology experts to retain only concepts specifically related to dosing errors. The resulting set of curated terms was then matched against the reported adverse events of each clinical trial.

For each study, dosing-related events were aggregated across all trial arms to estimate a trial-level dosing error rate based on the number of affected participants relative to the population at risk. A trial was labeled as positive when the lower bound of the 95% Wilson confidence interval of this rate exceeded a predefined threshold. Given the rarity of dosing errors and the pharmacovigilance practice of treating low-frequency events as potential safety signals (Beninger, 2020), a conservative threshold of 0.01% was adopted.

Using this procedure, approximately 4.6% of trials in the dataset were labeled as positive, resulting in a highly imbalanced classification problem that reflects the low prevalence of dosing errors in clinical research. A detailed description of the labeling workflow is provided in the companion dataset paper (Hêche et al., 2026).

The dataset was partitioned into training, validation, and test splits using a chronological strategy based on trial completion dates. This approach mitigates duration-related selection bias and provides a realistic estimate of prospective model performance. For the shared task, participants were provided with labeled training data and a validation set with masked labels during Phase 1, followed by an unlabeled test set for final evaluation in Phase 2. Table 5 summarizes key statistics of the dataset.

Field	Availability (%)	Mean chars	Median chars	Median words
allocation	99.62%	8.46	10	1
armDescriptions	99.40%	388.24	223	37
briefSummary	100%	502.40	338	50
conditions	100%	39.21	23	3
conditionsKeywords	63.94%	81.17	59	7
detailedDescription	61.93%	1,610.83	1,151	175
interventionDescriptions	100%	286.18	177	27
interventionModel	99.67%	9.34	8	1
interventionNames	100%	82.72	63	8
locationDetails	94.38%	1,310.26	123	22
protocolPdfText	42.33%	223,247.14	186,043	26,122

Table 2: Summary statistics of textual fields.

Feature	Avail.	#Cat.	Categories
healthyVolunteers	99.93%	2	FALSE, TRUE
masking	99.82%	5	NONE, SINGLE, DOUBLE, TRIPLE, QUADRUPLE
oversightHasDmc	87.82%	2	FALSE, TRUE
primaryPurpose	99.06%	10	TREATMENT, PREVENTION, DIAGNOSTIC, ECT, SUPPORTIVE_CARE, SCREENING, HEALTH_SERVICES_RESEARCH, BASIC_SCIENCE, DEVICE_FEASIBILITY, OTHER
sex	99.98%	3	ALL, FEMALE, MALE

Table 3: Features with categorical values.

Feature	Avail.	#Cat.	Categories
armGroupTypes	99.40%	6	EXPERIMENTAL, ACTIVE_COMPARATOR, PLACEBO_COMPARATOR, SHAM_COMPARATOR, NO_INTERVENTION, OTHER
interventionTypes	100%	11	DRUG, DEVICE, BIOLOGICAL, PROCEDURE, RADIATION, BEHAVIORAL, GENETIC, DIETARY_SUPPLEMENT, COMBINATION_PRODUCT, DIAGNOSTIC_TEST, OTHER
phases	99.99%	6	NA, EARLY_PHASE1, PHASE1, PHASE2, PHASE3, PHASE4

Table 4: Features with list of categorical values.

3. CT-DEB’26

The Clinical Trial Dosing Error Benchmark 2026 (CT-DEB’26) was organized as an open shared task with the goal of benchmarking ML approaches for predicting dosing error risk in interventional clinical trials. The challenge was hosted on the CodaBench platform² and is part of the CL4Health workshop at LREC 2026. Emphasis was placed on reproducibility, transparency of methodology, and realistic evaluation of prospective trial-level risk prediction.

3.1. Task Definition

Participants were tasked with predicting, for each clinical trial protocol in the dataset, the probabil-

²<https://www.codabench.org/>

Statistic	Value
Total trials	42,112
Training set	29,478
Validation set	6,316
Test set	6,318
Positive label prevalence	4.62%
Trials with protocol PDF text	42%

Table 5: Overview of the CT-DEB dataset.

ity that the trial would exhibit an elevated rate of dosing errors during its execution. Formally, given a feature vector $\mathbf{x}_i$ describing trial i , models were required to estimate a scalar probability $\hat{y}_i \in [0, 1]$ corresponding to the likelihood of the positive class, whereas the prediction target was a binary label as described in Section 2.

Submissions consisted exclusively of trial-level probability estimates, one per trial, identified by the unique ClinicalTrials.gov identifier (`nctid`). Participants were free to use any modeling approach, including classical ML, deep learning, or hybrid methods, provided that predictions were generated locally and only inference outputs were submitted for evaluation.

3.2. Challenge Phases

CT-DEB’26 was conducted in two sequential phases. During Phase 1 (validation phase), participants were provided with labeled training data and a validation set in which the target labels were masked. This phase was intended to support model development, hyperparameter tuning, and exploratory analysis, while preventing leaderboard overfitting. Fifteen teams submitted valid predictions during this phase.

Phase 2 (test phase) evaluated generalization performance on a held-out test set with no released labels. Leaderboards remained hidden during the active phase and were revealed only after the submission deadline. Four teams submitted final predictions during Phase 2, two of which (AL-Smadi, 2026; Hamnett et al., 2026) subsequently

contributed full technical papers to the CL4Health workshop.

3.3. Evaluation Protocol

Submitted predictions were evaluated by comparing the estimated probabilities against the ground truth labels on the hidden validation or test sets. The primary evaluation metric used for leaderboard ranking was the area under the receiver operating characteristic curve (ROC-AUC), which evaluates the ability of a model to rank positive instances above negative ones independently of a fixed decision threshold. Given the strong class imbalance of the dataset, additional metrics such as sensitivity, specificity, macro- F_1 , balanced accuracy, and Brier score were also reported to provide complementary perspectives on model performance.

In addition to ROC-AUC, several secondary metrics were reported for descriptive purposes, including macro-averaged F_1 score, sensitivity, specificity, and balanced accuracy. These metrics provide complementary insights into model behavior under different operating points, particularly with respect to minority-class detection.

All submissions were executed locally by participants, and only prediction files in a standardized CSV format were uploaded to the platform. No training or inference was performed on the CodaBench servers.

4. Results

4.1. Baseline

An official baseline was established to provide a reference point for the CT-DEB'26 shared task and to contextualize participant submissions. The baseline model is based on a CatBoost classifier (Dorogush et al., 2018) trained on the structured components of the dataset, with explicit handling of heterogeneous feature types derived from the dataset schema.

Schema-aware feature grouping: The dataset released for the shared task exposes a structured schema through the Hugging Face dataset interface, which specifies the semantic type of each feature (e.g., numerical value, categorical label, or sequence of labels). The baseline pipeline automatically parses this schema and partitions the input variables into four groups: (i) numerical features (e.g., enrollment counts, number of arms, and number of interventions), (ii) scalar categorical features corresponding to controlled vocabularies (e.g., primary purpose, masking strategy, or participant sex), (iii) list-valued categorical features representing multi-label protocol attributes (e.g., trial

phases, intervention types, or arm group types), and (iv) free-text fields extracted from registry entries and protocol documents. This automatic routing ensures that each feature type is processed using a transformation compatible with its underlying structure.

Feature preprocessing. Numerical variables are used directly as scalar inputs. Scalar categorical variables are encoded using one-hot representations derived from their predefined vocabularies. List-valued categorical variables are transformed using multi-hot encoding, where each possible category is represented by a binary indicator reflecting its presence within a given trial. Textual fields are intentionally processed using a minimal representation rather than deep linguistic models: each string-valued field is converted into a single numeric feature corresponding to its character length. This applies both to long-form narrative fields (e.g., trial summaries or protocol descriptions) and shorter string-valued registry attributes. This design choice keeps the baseline lightweight while emphasizing the predictive signal contained in structured trial metadata rather than relying on semantic modeling of protocol text.

Model training: The resulting tabular feature matrix is used to train a CatBoost classifier (Dorogush et al., 2018), a gradient-boosted decision tree model designed for heterogeneous tabular data. CatBoost was selected because of its strong empirical performance on structured datasets and its ability to handle categorical features efficiently. The model is trained on the official training split, with predictions expressed as probabilities for the positive class (elevated dosing error risk). These probabilities are used directly for evaluation using the ROC-AUC metric specified by the shared task.

The baseline therefore establishes a strong tabular reference point against which more complex multimodal approaches incorporating deep text representations can be evaluated.

Baseline performance: Despite its relatively simple design, the CatBoost baseline achieved strong performance and ranked first in both phases of the challenge. On the Phase 1 (validation) leaderboard, the baseline obtained an ROC-AUC of 0.8606, with a macro- F_1 score of 0.6291 and a balanced accuracy of 0.7062. On the Phase 2 (test) leaderboard, it achieved an ROC-AUC of 0.8624, with a macro- F_1 score of 0.6276 and a balanced accuracy of 0.7291. These results indicate that a schema-aware tabular learning approach already captures a substantial portion of the signal relevant for trial-level dosing error risk prediction, highlighting the importance of structured trial design variables.

4.2. Submissions

Participation overview: The shared task attracted 15 official submissions during Phase 1 and 4 submissions during Phase 2.

Leaderboard summary: In Phase 1, several submissions achieved ROC-AUC scores close to that of the baseline, with the best non-baseline entry reaching 0.8599.

In Phase 2, the highest-ranking non-baseline submission achieved an ROC-AUC of 0.8466, followed by a submission with a ROC-AUC of 0.8231.

Tables 6 and 7 present the official leaderboards for Phases 1 and 2, respectively, sorted by ROC-AUC.

Submissions with accompanying papers: Two submissions were accompanied by technical papers presented at the CL4Health workshop, offering complementary methodological perspectives.

ALSmadi (AL-Smadi, 2026): This submission adopts a feature-engineering-driven approach centered on gradient-boosted decision trees (LightGBM) trained on a high-dimensional representation of trial text. The authors construct a 3,451-dimensional feature space combining: (i) sparse lexical features (TF-IDF word unigrams and character n -grams), (ii) dense sentence embeddings (all-MiniLM-L6-v2) (Reimers and Gurevych, 2019; Wang et al., 2020), (iii) domain-inspired pattern features capturing dosing units, routes of administration, dosing frequency expressions, dose adjustments, and related cues, and (iv) scalar probability scores produced by fine-tuned transformer classifiers (BiomedBERT (Gu et al., 2021) and DeBERTa-v3 (He et al., 2021)). The model is trained using a stratified 5-fold ensemble with class-weighted loss to address the severe class imbalance of the dataset. Through systematic ablation experiments, the authors show that sentence embeddings contribute the largest individual performance gain, while sparse lexical features remain strongly complementary despite the presence of modern embedding methods. Their analysis further reveals that feature selection can improve predictive performance: restricting the model to the top 500–1000 most informative features yields slightly higher cross-validation ROC-AUC than using the full feature space, suggesting that aggressive feature selection acts as a regularization mechanism in the high-dimensional setting. Excluding the baseline, the corresponding participant ranked first in both phases (ROC-AUC: 0.8599 in Phase 1; 0.8466 in Phase 2).

Hamnett et al. (Hamnett et al., 2026): This submission investigates domain-specific transformer encoders as the primary mechanism for

representing protocol text and compares several biomedical language models, including ClinicalBERT (Alsentzer et al., 2019), PubMedBERT (Gu et al., 2021), BioBERT (Lee et al., 2020), and MedCPT (Jin et al., 2023). The authors embed multiple CT textual fields (e.g., brief summary, detailed description, arm and intervention descriptions) using mean-pooled transformer representations, then combine these embeddings with one-hot encoded categorical trial design variables (including intervention model, allocation, oversight, and phase). The resulting representations are evaluated using a range of downstream classifiers and neural architectures—including logistic regression and support vector classifiers, gradient-boosting models (XGBoost/LightGBM), and feed-forward networks with residual connections. The authors also explore the role of class weighting in mitigating the strong class imbalance present in the dataset. Across their experiments, BioBERT yielded the strongest text representations among the evaluated encoders, whereas concatenating embeddings from multiple transformer models did not provide additive gains in their experimental setting. Excluding the baseline, the corresponding submission ranked second in Phase 2 (ROC-AUC: 0.8231).

Methodological observations: Although the two submissions adopt different modeling paradigms, they share several common design choices. First, both approaches combine textual representations with structured trial metadata derived from ClinicalTrials.gov registry records.

Second, both studies confirm that classical ML models such as gradient boosting remain highly competitive for this task, achieving performance comparable to more complex neural architectures when trained on sufficiently informative representations. Finally, both systems emphasize the complementary roles of lexical and semantic text features: sparse textual cues (e.g., domain-specific vocabulary or dosing expressions) and dense contextual embeddings capture different aspects of protocol descriptions and together contribute to improved predictive performance. These findings illustrate the diversity of viable modeling strategies for the CT-DEB'26 task and suggest that hybrid approaches combining structured trial metadata with semantically rich text representations constitute a promising direction for future research.

Observed performance characteristics: Across both phases, submissions exhibited diverse operating behaviors, with some models favoring conservative predictions with very high specificity and others prioritizing sensitivity to detect a larger fraction of trials with elevated dosing error risk.

Submission	ROC-AUC	F1 (macro)	Sensitivity	Specificity	Bal. Acc.	Brier score	Rank
Baseline	0.8606	0.6291	0.4950	0.9174	0.7062	0.0747	-
(AL-Smadi, 2026)_475872	0.8599	0.5166	0.0301	0.9990	0.5146	0.0434	1
IDEAMCVG_459849	0.8561	0.5101	0.8930	0.6791	0.7860	0.1848	2
ahajighasem_449348	0.8561	0.5101	0.8930	0.6791	0.7860	0.1848	2
kzan25_460627	0.8561	0.5101	0.8930	0.6791	0.7860	0.1848	2
uccaeid_480109	0.8561	0.5101	0.8930	0.6791	0.7860	0.1848	2
mshamani_459802	0.8555	0.5204	0.8528	0.7037	0.7783	0.1885	6
5Gmodels_459893	0.8495	0.5076	0.8729	0.6789	0.7759	0.2001	7
MLtests_465716	0.8490	0.5425	0.7960	0.7497	0.7728	0.1629	8
msha25_465724	0.8489	0.5206	0.8294	0.7088	0.7691	0.1791	9
TCVG_465640	0.8460	0.5362	0.8227	0.7346	0.7787	0.1875	10
alha35_465681	0.8460	0.5362	0.8227	0.7346	0.7787	0.1875	10
mehla40_465693	0.8460	0.5362	0.8227	0.7346	0.7787	0.1875	10
IAU CV team_465636	0.8460	0.5362	0.8227	0.7346	0.7787	0.1875	10
farhankhan_462617	0.8426	0.5157	0.7960	0.7078	0.7519	0.1796	14
yanh_h_zry_474851	0.5000	0.4879	0.0000	1.0000	0.5000	0.0473	15

Table 6: Official Phase 1 (validation) leaderboard for CT-DEB’26, sorted by ROC-AUC.

Submission	ROC-AUC	F1 (macro)	Sensitivity	Specificity	Bal. Acc.	Brier score	Rank
Baseline	0.8624	0.6276	0.5676	0.8906	0.7291	0.0865	-
(AL-Smadi, 2026)_530456	0.8466	0.5072	0.0235	0.9963	0.5099	0.0466	1
(Hamnett et al., 2026)_510369	0.8231	0.5010	0.8588	0.6529	0.7559	0.1838	2
MLtests_511064	0.5000	0.4862	0.0000	1.0000	0.5000	0.0538	3
5Gmodels_511131	0.5000	0.4862	0.0000	1.0000	0.5000	0.0538	3

Table 7: Official Phase 2 (test) leaderboard for CT-DEB’26, sorted by ROC-AUC.

This diversity is visible in the leaderboard. For example, the (AL-Smadi, 2026) submission achieves extremely high specificity but very low sensitivity in both phases (Phase 1 sensitivity: 0.0301; Phase 2 sensitivity: 0.0235), indicating a highly conservative strategy that flags very few trials as high risk. In contrast, the (Hamnett et al., 2026) submission favors sensitivity (Phase 2 sensitivity: 0.8588) at the expense of specificity (0.6529), reflecting a more aggressive detection strategy that identifies most positive trials while producing more false positives. While ROC-AUC served as the primary ranking metric, the additional evaluation measures provide complementary perspectives on model behavior under different operating points. Overall, the results illustrate that CT-DEB’26 supports meaningful comparison of modeling strategies ranging from schema-aware tabular baselines to text-centric and hybrid approaches.

4.3. Discussion and Lessons Learned

The results of the CT-DEB’26 shared task highlight several important observations regarding the prediction of dosing errors from clinical trial protocols.

First, the strong performance of the CatBoost baseline demonstrates that a substantial portion of the predictive signal is already captured by structured trial metadata and high-level protocol descrip-

tors. Despite relying only on simple statistics derived from textual fields and schema-aware processing of tabular variables, the baseline achieved the highest ROC-AUC in both phases. This suggests that features related to trial design (e.g., intervention types, masking strategy, study phase, and enrollment characteristics) contain meaningful information about the operational complexity of a trial and therefore about the potential risk of dosing errors.

Second, the submissions illustrate complementary modeling paradigms for leveraging the rich textual information contained in trial documentation. The feature-engineering approach of ALSmadi (AL-Smadi, 2026) demonstrates that carefully constructed lexical and semantic representations combined with gradient-boosted models can effectively capture dosing-related patterns in protocol descriptions. In contrast, the work of Hamnett et al. (Hamnett et al., 2026) emphasizes the role of biomedical transformer models for producing contextualized text embeddings. Despite these methodological differences, both approaches confirm that hybrid models combining structured trial variables with textual representations provide the most promising direction for this task.

Third, the leaderboard results highlight the importance of considering multiple evaluation metrics when analyzing system behavior under strong

class imbalance. While ROC-AUC served as the primary ranking metric because it evaluates the ability of models to discriminate between positive and negative trials independently of a fixed decision threshold, the additional metrics provide important complementary perspectives. For example, some submissions achieved high ROC-AUC values while exhibiting extremely low sensitivity, indicating highly conservative prediction strategies that identify only a small fraction of trials with elevated dosing error risk. Conversely, other systems achieved higher sensitivity but substantially lower specificity, leading to larger numbers of false positive predictions. The Brier score further provides insight into the calibration quality of probabilistic predictions by measuring how closely predicted probabilities align with observed outcomes. Such calibration properties may be particularly relevant for potential downstream use of these models in risk assessment or decision-support settings.

Despite these encouraging results, several limitations of the current benchmark should be acknowledged. The dataset is derived from publicly available ClinicalTrials.gov registry entries and associated protocol documents, which provide only a partial view of the full operational complexity of clinical trials. In practice, dosing errors may arise from detailed procedural instructions, investigator decisions, or site-level deviations that are not fully captured in registry metadata. Moreover, the binary formulation of the task collapses potentially heterogeneous types of dosing errors into a single label, whereas real-world scenarios may require distinguishing between different categories of protocol deviations (e.g., incorrect dose amount, timing deviations, or administration route errors). As a result, the benchmark should be viewed primarily as a proxy task for studying trial-level risk signals rather than as a complete operational detection system.

Another challenge concerns the difficulty of interpreting model predictions in a regulatory or clinical quality assurance context. While gradient-boosting models allow feature importance analysis, the influence of textual features remains difficult to interpret at a semantic level. Developing explainable approaches that identify specific protocol elements or linguistic patterns associated with dosing error risk would be an important step toward practical deployment.

These observations suggest several directions for future research. One promising avenue is the application of more advanced natural language processing techniques to extract structured dosing instructions directly from full protocol documents. For example, information extraction systems could identify dosage amounts, schedules, administration routes, and protocol modification rules from long-form protocol PDFs with non-uniform structures,

enabling more fine-grained modeling of dosing complexity. Another direction involves the systematic study of feature importance and model explainability, including techniques such as SHAP analysis (Lundberg and Lee, 2017) or attention-based attribution methods, to better understand which aspects of trial design contribute most strongly to predicted risk. Finally, future datasets could incorporate richer supervision signals, including multi-class taxonomies of medication errors or annotations of specific protocol deviations, enabling models that move beyond binary risk prediction toward more actionable clinical insights.

Overall, the CT-DEB'26 shared task provides a first benchmark for studying dosing error prediction from clinical trial protocols and highlights both the promise and the challenges of applying ML to this safety-critical problem.

5. Conclusions

This paper presented an overview of the CT-DEB'26 shared task on predicting dosing errors in interventional clinical trials. The task introduced a new benchmark dataset derived from ClinicalTrials.gov that combines structured trial design variables with long-form textual protocol descriptions. By framing dosing error prediction as a ML problem, the challenge aimed to stimulate research at the intersection of clinical trial methodology, medication safety, and natural language processing.

The shared task attracted a diverse set of modeling approaches. While the official baseline relied on a schema-aware tabular learning pipeline using structured trial metadata, participating teams explored complementary strategies incorporating richer textual representations, including large-scale feature engineering pipelines and transformer-based biomedical language models. The results demonstrate that structured trial design features already provide substantial predictive signal for identifying trials at risk of dosing errors, while textual representations can provide complementary information about protocol details and dosing instructions.

Another important direction concerns model interpretability and explainability. In order for predictive models to be used in clinical trial planning or quality assurance settings, it is essential to understand which aspects of trial design contribute most strongly to predicted dosing error risk. Techniques for feature attribution and explainable ML may therefore play a key role in translating predictive models into actionable insights for trial designers and regulatory stakeholders.

Finally, advances in natural language processing offer opportunities to extend this benchmark by extracting more detailed dosing-related infor-

mation from full protocol documents. Methods for structured information extraction from long protocol PDFs could enable richer representations of dosing schedules, administration rules, and protocol modifications. Such developments would move beyond trial-level risk prediction toward automated identification of potential dosing issues during the early design stages of clinical studies.

Overall, CT-DEB'26 represents a step toward the development of data-driven methods for improving medication safety in clinical research. We hope that the dataset, baseline models, and shared task results will provide a useful foundation for future work on ML methods supporting safer and more efficient clinical trial design.

6. Acknowledgments

This work was supported by Innosuisse - the Swiss Innovation Agency - grant number 114.721 IP-ICT.

7. Bibliographical References

- Mohammad AL-Smadi. 2026. Automated detection of dosing errors in clinical trial narratives: A multi-modal feature engineering approach with lightgbm. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma de Mallorca, Spain. European Language Resources Association (ELRA).
- Emily Alsentzer, John Murphy, William Boag, Weihung 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.
- Jeffrey K Aronson. 2009. [Medication errors: definitions and classification](#). *British journal of clinical pharmacology*, 67(6):599–604.
- Paul Beninger. 2020. Signal management in pharmacovigilance: a review of activities and case studies. *Clinical therapeutics*, 42(6):1110–1129.
- Andrea Cipriani and Corrado Barbui. 2010. What is a clinical trial protocol? *Epidemiology and Psychiatric Sciences*, 19(2):116–117.
- Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.
- Rachel Ann Elliott, Elizabeth Camacho, Dina Jankovic, Mark J Sculpher, and Rita Faria. 2021. [Economic analysis of the prevalence and clinical and economic burden of medication error in england](#). *BMJ Quality & Safety*, 30(2):96–105.
- European Medicines Agency. 2023. Guideline for the notification of serious breaches of regulation (eu) no 536/2014 or the clinical trial protocol.
- European Parliament and Council of the European Union. 2014. [Regulation \(eu\) no 536/2014 of the european parliament and of the council of 16 april 2014 on clinical trials on medicinal products for human use, and repealing directive 2001/20/ec](#). Official Journal of the European Union, L 158/1.
- Sohrab Ferdowsi, Nikolay Borissov, Julien Knafou, Poorya Amini, and Douglas Teodoro. 2021. [Classification of hierarchical text using geometric deep learning: the case of clinical trials corpus](#). *arXiv preprint arXiv:2110.15710*.
- Sohrab Ferdowsi, Julien Knafou, Nikolay Borissov, David Vicente Alvarez, Rahul Mishra, Poorya Amini, and Douglas Teodoro. 2023. [Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study](#). *Patterns*, 4(3).
- 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 *ACL*.
- Leon Hamnett, Favour Igwezeke, Joseph Itopa, and Mary Adetutu Adewunmi. 2026. Detecting dosing errors in clinical trials using domain-specific transformer embeddings and classification models. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma de Mallorca, Spain. European Language Resources Association (ELRA).
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced bert with disentangled attention. In *ICLR*.
- Félicien Hêche, Sohrab Ferdowsi, Anthony Yazdani, Sara Sansaloni-Pastor, and Douglas Teodoro. 2026. Early risk stratification of dosing errors in clinical trials using machine learning. *arXiv preprint arXiv:2602.22285*.
- Alexander Hodkinson, Natasha Tyler, Darren M Ashcroft, Richard N Keers, Kanza Khan, Denham Phipps, Aseel Abuzour, Peter Bower, Anthony Avery, Stephen Campbell, et al. 2020. [Preventable medication harm across health care settings: a systematic review and meta-analysis](#). *BMC medicine*, 18:1–13.

- Félicien Hêche, Sohrab Ferdowsi, Anthony Yazdani, et al. 2026. [Machine learning for medication error detection: A scoping review](#). *Research Square*. Research Square preprint, Version 1. Available at: <https://doi.org/10.21203/rs.3.rs-8919709/v1>.
- International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). 2016. [Integrated addendum to ich e6\(r1\): Guideline for good clinical practice e6\(r2\)](#). ICH Harmonised Guideline.
- Qiao Jin, Xin Yuan, Sheng Wang, Yifan Zhang, Jiayuan He, and Zhiyuan Liu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. In *ACL*.
- 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.
- 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.
- Linda Martin, Melissa Hutchens, and Conrad Hawkins. 2017. Trial watch: clinical trial cycle times continue to increase despite industry efforts. *Nature Reviews Drug Discovery*, 16(3):157–158.
- MedDRA Maintenance and Support Services Organization. 2025. Medical dictionary for regulatory activities (meddra). <https://www.meddra.org/>. Accessed: 2025-09-12.
- Andrew Mulcahy, Stephanie Rennane, Daniel Schwam, Reid Dickerson, Lawrence Baker, and Kanaka Shetty. 2025. Use of clinical trial characteristics to estimate costs of new drug development. *JAMA Network Open*, 8(1):e2453275.
- National Library of Medicine (US). 2000. [Clinicaltrials.gov](#). Bethesda (MD): National Institutes of Health (US).
- Pydantic AI. Data validation using python type hints. <https://github.com/pydantic/pydantic>.
- Python Software Foundation. enum — support for enumerations. <https://docs.python.org/3/library/enum.html>. Accessed: 2026-01-07.
- Nils Reimers and Iryna Gurevych. 2019. Sentencebert: Sentence embeddings using siamese bert-networks. In *EMNLP*.
- Mahsa Rouhani, Maryam Mousavi, Mohammad Mahdi Kooshyar, Soodabeh Shahidsales, Yasha Makhdoumi, Amir Amirabadi, and Sepideh Elyasi. 2018. Application of a chemotherapy standard form in patients with breast cancer: comparison of private and public centers. *Jundishapur Journal of Natural Pharmaceutical Products*, 13(3).
- Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. 2022. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 12(7):3049–3062.
- Rayhan A Tariq, Rishik Vashisht, Ankur Sinha, and Yevgeniya Scherbak. 2018. Medication dispensing errors and prevention.
- Rayhan A. Tariq, Rishik Vashisht, Ankur Sinha, and Yevgeniya Scherbak. 2025. [Medication Dispensing Errors and Prevention](#). StatPearls Publishing, Treasure Island (FL).
- Douglas Teodoro, Nona Naderi, Anthony Yazdani, Boya Zhang, and Alban Bornet. 2025. [A scoping review of artificial intelligence applications in clinical trial risk assessment](#). *medRxiv*, pages 2025–01.
- U.S. Food and Drug Administration. 2024. [Title 21, code of federal regulations, part 312: Investigational new drug application](#). Code of Federal Regulations. Accessed via the Electronic Code of Federal Regulations (eCFR).
- Bernard Vrijens, Antoine Pironet, and Eric Tousse. 2024. [The importance of assessing drug exposure and medication adherence in evaluating investigational medications: ensuring validity and reliability of clinical trial results](#). *Pharmaceutical Medicine*, 38(1):9–18.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*.
- Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. 2018. [Estimation of clinical trial success rates and related parameters](#). *Biostatistics*, 20(2):273–286.
- Olivier J Wouters, Martin McKee, and Jeroen Luyten. 2020. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama*, 323(9):844–853.

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.

Anthony Yazdani, Alban Bornet, Philipp Khlebnikov, Boya Zhang, Hossein Rouhizadeh, Poorya Amini, and Douglas Teodoro. 2025. An evaluation benchmark for adverse drug event prediction from clinical trial results. *Scientific Data*, 12(1):424.

Overview of the CRF 2026 Shared Task on Clinical Case Report Forms Filling

Pietro Ferrazzi^{1,2}, Soumitra Ghosh¹, Alberto Lavelli¹, Bernardo Magnini¹

¹Fondazione Bruno Kessler, Povo, Trento, Italy

²University of Padova, Padova, Italy

Abstract

Case Report Forms (CRFs) are structured instruments widely used in clinical research to systematically collect patient information according to predefined protocols. In practice, CRFs are often manually completed by clinicians based on patients' clinical reports, a process that is time-consuming and prone to inconsistencies. Despite their central role in medical studies, automatic population of CRFs from clinical narratives remains largely underexplored in the Natural Language Processing community, partly due to the scarcity of publicly available datasets. In this paper, we present the CRF Filling Shared Task, organised at the CL4Health Workshop at LREC 2026, which aims to advance research on automatic extraction of structured clinical information from unstructured patient notes. The task consists of assigning the correct value to a set of predefined CRF items given a clinical note. The target dataset is derived from a real-world CRF for dyspnea assessment, comprising 134 medical items with predefined value sets. The task is provided in two languages, Italian and English. We describe the dataset, the task formulation, and the evaluation framework, and discuss the participating systems and their results. By introducing this shared task, we aim to stimulate research on clinically applicable NLP systems for structured data extraction in healthcare.

Keywords: case report form, shared task, medical information extraction, medical question answering

1. Introduction

Case Report Forms (CRFs) are structured data collection instruments used in clinical research to systematically record patient information. They are a fundamental tool in medical practice and research: clinicians use structured forms in Emergency Departments during patient admission, physicians rely on them to monitor specific conditions, and researchers use them to identify and characterize eligible patients for clinical studies (Richesson and Nadkarni, 2011; Bellary et al., 2014). By standardising how information is collected and represented, CRFs improve efficiency through normalisation and structured documentation, and enhance effectiveness by providing a clear, organised overview of available clinical data (Pétavy et al., 2019; Rinaldi et al., 2025; Lin et al., 2015). With the rapid progress of Natural Language Processing (NLP) methods, interest in automatic CRF filling has grown, as researchers explore whether automating this time-consuming task can help streamline clinical data collection and support medical research workflows (Mac Kenzie et al., 2016). Gutiérrez-Sacristán et al. (2024) were among the first to propose an automatic population of case report forms from clinical notes based on NLP techniques, while Crema et al. (2024) introduced the usage of Large Language Models for the task, by designing a BERT-based approach that frames the task as extractive question answering. These studies have opened a promising research direction with significant potential impact in real-world clinical settings, where substantial time is spent manually complet-

ing CRFs based on patients' clinical notes. However, the applied methods do not explore the full set of state-of-the-art NLP approaches, which remain largely underexplored. Furthermore, the mentioned approaches do not release any dataset that can be used to advance research in the field. For these reasons, we propose the CRF Filling Shared Task, hosted at the CL4Health Workshop at LREC 2026, to foster the development of systems capable of automatically populating CRFs from clinical notes (an intuitive example of the task is shown in Figure 1). We combine a set of data sources, spanning from manually annotated CRFs from emergency department settings (Kaczmarek et al., 2026) to semi-automatically annotated examples from scientific literature and medical exams (Ferrazzi et al., 2025), and more than two thousand manually selected unannotated clinical notes from those presented in Ferrazzi et al. (2026). All datasets are released through our HuggingFace organization account¹. In this paper, we present an overview of the Shared Task task, including the provided datasets, a brief description of the best systems proposed by each of the 12 participating teams, selected from the 32 valid test submissions. Moreover, we discuss the results by comparing submissions and identifying observed trends and similarities.

¹<https://huggingface.co/collections/NLP-FBK/crf-filling-sharedtask-cl4health2026>

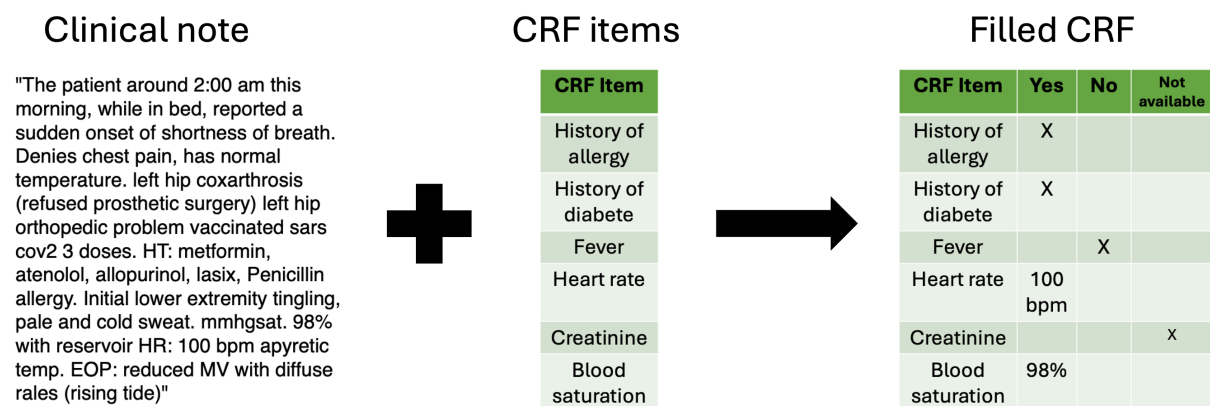


Figure 1: Example of the CRF filling task. Given a clinical note describing a patient’s history and medical conditions, the task is to automatically fill a predefined list of medical items based on the note’s content.

2. Task Description

The Dyspnea CRF consists of a list of 134 medical items that can assume different values. The task consists of assigning the correct value to each item given a patient’s clinical note.

As an example, the item *"chronic pulmonary disease"* has to be filled with a value from *"certainly chronic"*, *"certainly not chronic"*, *"possibly chronic"*. Whenever the note does not provide sufficient evidence, the correct output is *"unknown"*, which is always considered a valid answer.

The task is available in two languages: Italian and English. Multilingualism is intended to encourage the development of robust methods applicable across languages and diverse clinical settings, while also reflecting the linguistic variability of real-world healthcare data.

Participants are provided with an HuggingFace collection with all the data², the CodaBench portal for systems’ results submission³, and a GitHub repository for local evaluation during development⁴.

3. Datasets

The development of systems to automatically perform CRF filling is hampered by a lack of annotated data. This happens because manual annotation of clinical notes requires expert medical doctors, who are not often available for annotation tasks, given the workload in healthcare settings. Nevertheless, other sources can be leveraged to design systems that perform the task. For this reason, the

data we propose as training set combine (i) few gold-standard examples of the actual task at hand; (ii) much more unannotated data from the same data distribution, which experts did not annotate; and (iii) some annotated examples of tasks with some degrees of similarity with the task at hand, but not identical or coming from the same data distribution. Then, to train systems to fill the Dyspnea CRF, participants are provided with three datasets:

- 10 **gold-standard** pairs of clinical notes with manually filled Dyspnea CRF. These are intended to showcase the task. They can be used for few-shot settings, a base for data augmentation, etc.
- 2667 **unannotated** clinical notes about patients with Dyspnea. These are intended to provide participants with knowledge about the target data distribution.
- 71 (80 for Italian) pairs of clinical notes and **semi-automatically annotated** CRF. These are intended as examples of how the CRF filling task can be performed in domains other than Dyspnea.

The gold-standard pairs are the ones that describe the task at hand, while the unannotated clinical notes about Dyspnea patients and the semi-automatically annotated pairs are provided as extra data, potentially helpful in solving the task, but not strictly necessary. **Development** and **test** can be performed on two gold-standard datasets of 80 and 200 clinical notes, respectively, coming from the same source as the 10 provided for training (Kaczmarek et al., 2026).

Gold-standard The gold-standard dataset (Kaczmarek et al., 2026) is comprises (i) clinical notes reporting patients’ history and (ii) the manually annotated Dyspnea CRF, a list of 134 medical items

²<https://huggingface.co/collections/NLP-FBK/crf-filling-sharedtask-cl4health2026>

³<https://www.codabench.org/competitions/11984/>

⁴<https://github.com/hltfbk/CRF-filling-CL4Health2026>

filled by medical doctors, designed to capture clinical information in the setting of an Emergency Department (ED). For each clinical note, medical doctors from the eCREAM European Project have manually filled the 134 CRF items.

The *clinical notes* (10 for training, 80 for development, and 200 for testing) refer to patients of the San Giovanni Bosco hospital in Turin (Italy) who went to the emergency department with Dyspnea (a respiratory condition that involves shortness of breathing). All notes have been anonymized to preserve privacy according to the ethical protocol defined in the context of the eCREAM European Project. All information that could identify patients, such as family members, locations, names, and phone numbers has been replaced by placeholders. The notes come from different sources inside the Emergency Department, including anamnesis reports, triage records, nursing care notes, specialist consultation notes, medical visit reports, diagnostic reports, and discharge summaries.

Each *CRF item* is associated with a predefined label space that reflects the type of information to be extracted. If the note does not contain the information required to fill an item, its filling value is set to "unknown". The most common label types include: (i) binary categorical labels (e.g., yes/no, positive/negative, present/absent), (ii) ordinal labels, typically used to indicate whether a measurement falls below, within, or above the normal range (e.g., body temperature can be hypothermic, normothermic, hyperthermic), and (iii) for those clinical tests whose outcomes require contextual interpretation, the filling value can be "measured" if it has been collected, "unknown" otherwise.

The dataset is characterized by a **high imbalance towards "unknown"** values. More than 90% of the items are annotated with such label, as the CRF is designed to collect all the potentially relevant information, while patients' notes often report just a small fraction of them.

Gold-standard data distribution On average, each gold-standard note contains 5.7 annotated items, although the distribution is uneven: some notes include more than 15 annotated items, while around 20% (58/290) contain none. The average length of the notes is 131 words, with some outliers reaching up to 400 words. Note length also varies across document types, with triage and discharge records typically shorter than others.

Coverage of the CRF is partial, with approximately three-quarters of the items annotated at least once in the dataset. The annotations also exhibit characteristic properties of clinical documentation. In particular, redundancies are common, with 28% of notes containing repeated annotations of the same item with identical values, while a small

proportion (0.8%) includes multiple assignments of the same item with differing values. These patterns do not reflect annotation errors, but rather the longitudinal and sometimes evolving nature of clinical information.

Unannotated Notes We provide a set of 2667 unannotated clinical notes of patients with Dyspnea as additional source of knowledge about this medical condition. These notes are a manually selected subset from the much larger dataset of clinical cases presented in [Ferrazzi et al. \(2026\)](#), which contains 1.9M anonymized clinical notes from the emergency department of the San Giovanni Bosco hospital, Turin (Italy).

Semi-automatic CRFs As a further resource to enhance task understanding at training time, we include the semi-automatically generated CRFs proposed in [Ferrazzi et al. \(2025\)](#). There, we proposed a methodology to convert medical datasets curated for Named Entity Recognition into annotated CRF filling data by (i) creating shared lists of items for clinically similar notes, and (ii) automatically filling each item utilizing the available manual annotation information. We utilise the dataset composed of 71 Italian and 80 English notes. Such notes are grouped into clusters (7 for Italian, 8 for English) based on patients' histories and conditions. Each cluster of notes is associated with a cluster-specific CRF that contains medical items relevant to the cluster itself. For this reason, these items do not necessarily overlap with the target Dyspnea CRF ones, but are intended as a means to learn the CRF filling task, regardless of the medical characteristics.

Data Translation The clinical notes collected from the San Giovanni Bosco hospital are in Italian, while the semi-automatic ones are both in Italian and English. As we aim to propose a multilingual task which includes both Italian and English, we translated the clinical notes into English, following two different approaches. The *gold-standard* clinical notes have been manually translated into English by professional translators, while the unannotated clinical notes have been automatically translated into English using GPT-5. For the automatic approach, we evaluate the translation quality via back-translation ([Rapp, 2009](#)). Such a method involves translating the target text back into Italian and comparing the original version with the back-translated one. While this process may compound errors across translation steps, prior work has shown that back-translation scores correlate with human judgments and provide a practical proxy for translation quality ([Zhuo et al., 2023](#)). We compare the back-translated text with the source uti-

lizing four metrics, showcasing good translation quality (COMET, CHRf (Popović, 2015), CHRf++ (Popović, 2017), and BERTScore (Zhang et al., 2020)). We report the results in Table 1. The prompt utilized for automatic translation can be found in the Appendix.

Metric	Score
BERTScore	91.8
COMET	79.7
chrF	71.9
chrF++	68.3

Table 1: Back-translation evaluation metrics for the unannotated 2667 notes (Italian to English via GPT-5).

4. Evaluation

Model predictions are evaluated by comparing them to the ground truth label for each CRF item, using exact match as the evaluation criterion. The most intuitive metric for CRF filling would then be accuracy, calculated by comparing the filling value prediction to the ground truth for each item in the CRF among all clinical notes. This would quantify how many item values are filled correctly across the dataset. Nevertheless, the labels distribution is highly imbalanced among items, with more than 90% of them being labelled as "unknown". For this reason, a system that always predicts such a value would result in extremely high performances, which would not adequately represent reality. Therefore, we utilise **macro-F1** as the primary evaluation metric instead. This choice rewards systems that properly assign rare values, balancing the distribution-shift effect. For instance, a rare label is "*short*", which can only be assigned to two CRF items, "duration of patient's consciousness recovery" and "duration of patients' unconsciousness". A system that does not learn to correctly assign this label will be penalised as much as a system that consistently misses the "unknown" label. This choice allows to identify systems that are capable of identifying less frequent, clinically informative outcomes.

Baseline The baseline we select to tackle the task is an LLM (Qwen3-8B without thinking by Yang et al. (2025)) prompted via 2-shot. For each clinical note, we prompt the LLM as many times as CRF items exist (i.e., 134 times), providing (i) the clinical note, (ii) the CRF item, (iii) two examples. The examples are selected from the training set, and always refer to the same item as the target one. The prompt is presented in Appendix. The baseline achieves a macro-F1 of 47.0 and 53.9 in the English and Italian test sets, respectively,

5. Task Results

In this section, we report the performances of the various participating systems on the task. Table 2 shows the results for the best valid test submissions for English, while Table 3 for Italian. We present the results highlighting whether a system uses any closed-source LLM or not. While this does not represent a limitation of any kind with respect to the shared task, we believe it is useful to understand how submissions that only rely on open-source systems behave with respect to closed-source ones, as they can be locally deployed in healthcare institutions.

The team that achieved the best performance on both Italian and English is *Aurum*, who divides the 134 items into 14 groups, and optimizes a Chain-of-Thought prompt for each group using Qwen3-Max. The best system relying on open-source models for English is *MdL*, which leverages an entity recognition step before prompting Qwen3-8B via 1-shot. For Italian, the best performing system with open-source models was proposed by the *Polimi* team, which divided the items into three groups based on the options labels, and prompt Mistral-small-3.2-24B-Instruct with a prompt including a glossary of abbreviations, followed by impactful postprocessing.

In Table 4 we report the comparison of the results on Italian and English for those submissions that handled both languages. It can be seen that systems perform better in English.

Team name	CS	Macro-F1
Aurum	✓	68.3
DocUA	✓	62.8
Gladiators	✓	59.7
MdL		57.1
NM	✓	55.7
SEBIS		55.3
DocUA		54.4
Jovian_Tech	✓	52.4
Greyc		48.1
Innov8rs	✓	47.1
BASELINE		47.0
LTRC-CRF		38.3

Table 2: Results for English test. "CS" refers to the system using any closed-source model.

6. Submissions

In this section, we provide an overview of the ten submissions received through the Test Phase of the CodaBench portal, together with a detailed description. We report a summary of some relevant

Team Name	CS	Macro-F1
Aurum	✓	67.2
Polimi		63.5
DocUA	✓	57.0
NM	✓	55.6
BASELINE		53.9
Innov8rs		44.7
LTRC-CRF		29.4

Table 3: Results for Italian test. "CS" refers to the system using any closed-source model.

Team Name	Macro-F1			
	IT	EN	δ	$\delta\%$
Aurum	67.2	68.3	-1.1	-2%
DocUA	57.0	62.8	-5.8	-9%
NM	55.6	55.7	-0.1	0%
Innov8rs	44.7	47.1	-2.4	-5%
LTRC-CRF	29.4	38.3	-8.9	-23%

Table 4: Side-by-side comparison of the Macro-F1 scores for the Italian and English datasets. The delta columns show the difference in scores and the percentage drop from English to Italian.

dimensions common among submissions in Table 5.

SEBIS SEBIS proposes a two-stage pipeline for CRF filling based on MedGemma-27B. Their approach separates the task into two steps: first, predicting the presence of the information associated with each CRF item, and then extracting the corresponding value when applicable. This decomposition is designed to reduce hallucinations and limit over-interpretation by enforcing closer grounding in the clinical text. The system relies on few-shot prompting without model fine-tuning.

Gladiators Gladiators proposes a hybrid system that combines Claude-4.6-Opus with rule-based post-processing. The LLM is used as the primary extractor with deterministic prompting (zero temperature), while the prompt design emphasises key extraction rules through repetition to increase the model’s attention to frequently occurring clinical variables. A set of post-processing rules addresses common extraction issues, including Unicode normalisation for medical subscripts, regular-expression-based extraction of vital signs, and explicit handling of symptom negation. Additional checks verify treatments based on the presence of action verbs to reduce false positives. The system adopts a conservative strategy by defaulting to *unknown* in the absence of explicit textual evidence.

Innov8rs Innov8rs adopts a large language model ensemble approach. For the English track, they combine Gemini 2.5 Flash and Llama 3.3 70B, using a dynamic few-shot prompting strategy in which in-context examples are selected based on TF-IDF similarity to the input note. For the Italian track, the system relies on Llama 3.3 70B with static few-shot examples derived from the gold training data. To mitigate hallucinations, the authors introduce a precision-oriented filtering mechanism calibrated on the development set, which suppresses predictions for CRF items with historically high false-positive rates. This strategy is designed to improve overall macro-F1 performance by favoring precision on difficult items. Nevertheless, they experience that this filtering does not generalize to the test set.

Jovian_Tech Jovian_tech employs a few-shot prompting approach based on Claude Haiku 4.5 to automatically populate Dyspnea CRFs from clinical notes. Their system uses three gold-standard examples as in-context demonstrations to guide the extraction of the 134 CRF variables. The prompt design emphasizes conservative predictions and strict adherence to the predefined value sets for each item. The pipeline includes text normalization of the input notes, LLM-based extraction, and validation of the predicted values against the allowed option sets. The approach favors *unknown* predictions in the absence of explicit textual evidence, reflecting the sparsity of relevant information in clinical reports.

NM NM proposes a two-stage pipeline combining extraction and verification. In the first stage, a few-shot prompted Gemini 3 Flash Preview model populates CRF fields directly from the clinical note, prioritizing recall but generating a number of false positives. In the second stage, LLaMA4 Maverick 17B model reviews both the original note and the preliminary CRF output to verify each prediction and remove unsupported or ambiguous entries. This sequential strategy aims to balance coverage and precision by retaining only values that are clearly supported by the source text.

Aurum Aurum presents a modular extraction framework implemented with DSPy, decomposing the 134 CRF items into 14 domain-specific extractors (e.g., medical history, lab values, treatments, vital signs). Each module is defined through a typed DSPy signature with chain-of-thought reasoning, enabling targeted extraction within a specific clinical subdomain rather than relying on a single monolithic prompt. The prompts were iteratively refined through error analysis to reduce systematic false positives and negatives. The system uses Qwen3-

Rank	Team	Grouped	CS	Pipeline	FP/FN trade-off	Few-shot	Post-proc	Extra data
1	Aurum	✓	✓	✓	✓			
2	Polimi			✓	✓		✓	✓
3	DocUA	✓	✓	✓	✓		✓	
4	Gladiators	✓	✓	✓			✓	
5	MdL	✓		✓			✓	
6	NM		✓	✓	✓	✓		
7	SEBIS			✓	✓	✓		
8	Jovian_tech	✓	✓	✓		✓		
9	Greyc			✓	✓	✓		
10	Innov8ors		✓	✓	✓	✓	✓	

Table 5: Revised overview of system design choices across participating teams. Checkmarks indicate only explicitly supported evidence from system descriptions. Blank cells denote insufficient or implicit evidence. **Grouped** refers to the fact that multiple items have been grouped together into one single prompt; **CS** to the usage of closed-source models in the submission; **Pipeline** to the concatenation of multiple steps (instead of a single LLM call); **FP/FN trade-off** to whether the submission analyses and tries to optimize the trade-off between generating as few False Positives as possible, at the cost of generating more False Negatives; **Few-shot** if at any step, few shots are provided as examples; **Post-proc** refers to the presence of deterministic post-processing on the predictions; **Extra data** for the use of a dataset other than the gold-standard.

Max (Thinking) as the underlying model and performs consistently across both languages.

MdL MdL adopts a two-step approach combining rule-based detection with LLM-based normalization. First, relevant CRF item categories are identified through synonym matching and entity recognition using the GliNER model, with synonyms compiled from the training set to capture common variants and acronyms. Predictions from both methods are merged to select the most relevant item categories and filter out those likely to be *unknown*. In the second step, the selected items and the clinical note are provided to a Qwen 8B model through a structured prompt to normalize the extracted information and assign the appropriate CRF values according to predefined rules.

Polimi Polimi focuses on the Italian track. After evaluating several configurations, their best system uses Mistral-small-3.2-24B-Instruct in a zero-shot setting with three specialized prompts tailored to different item types (binary, categorical, and measured variables). The prompts incorporate a glossary of clinical abbreviations automatically extracted from the unlabeled notes. In addition, rule-based post-processing is applied to items with low recall but identifiable textual patterns. The system was tuned on the development set and tends to favor conservative predictions, producing more false negatives than false positives.

Greyc Greyc proposes a modular LLM-based framework consisting of three components: a rewriting model, an extractor, and a judging model. The rewriting stage reformulates the clinical note with respect to each CRF item, aiming to highlight relevant information. A second LLM then extracts candidate values using a k-shot prompting strategy, while a third LLM acts as a judge to verify whether sufficient textual evidence supports the prediction, defaulting to unknown otherwise. While the rewriting stage improves recall, it also increases false positives, which the judging step attempts to mitigate through a more conservative verification process.

DocUA DocUA introduces LLM StructCore, a two-stage system combining schema-guided reasoning with deterministic post-processing. In the first stage, Mistral Large 3 (675B) generates structured JSON summaries using a set of schema-guided reasoning patterns designed to ensure consistent output structure. In the second stage, deterministic rules perform canonicalization, vocabulary normalization, and mapping to standardized medical concepts through UMLS, along with evidence-based checks to reduce false positives. This hybrid design combines LLM-based extraction with rule-based normalization to produce structured CRF outputs.

7. Discussion

The participating systems reveal several emerging trends in approaches to automatic CRF fill-

ing from clinical notes. All submissions rely on large language models (LLMs), often combined with structured prompting strategies and multi-steps pipelines to improve reliability and reduce hallucinations. Almost all submissions identify as main challenge the high sparsity of the data, given that around 95% of the CRF items' labels are set to *unknown*, and analysed the trade-off between False Positives and False Negatives. Only one submission (Polimi) utilized data other than the gold-standard set, achieving the best results for Italian, and the best results overall without using closed-source models.

Conservative prediction strategies. Many systems adopt conservative strategies to avoid hallucinations, frequently defaulting to the *unknown* value unless clear textual evidence is present. Overall, participants observed that LLMs always tend to fill items with values other than *unknown* even when explicit information is not present in the clinical note. Many teams present the task as a trade-off between generating as few False Positives (FP) as possible, while accepting to generate some False Negatives (FN).

For instance, Innov8ors introduce a development-set calibrated filter that suppresses predictions for items with historically many FP, which results in too much bias towards the development set. NM reports that open-weight models tend to reduce FP, being more conservative than closed-source counterparts. Greyc show that their judging component eliminates a large portion of FP but also removes a substantial number of true positives. Aurum design works at a prompt level, achieving a 92% reduction of FP, at the cost of a 30% increase of FN. Polimi observed that LLM-based verification is not as effective as rule-based postprocessing driven by pattern finding in the original clinical notes.

Multi-stage pipelines. All teams decomposed the task into multiple stages. For example, SEBIS and NM employ a two-step approach, separating extraction from verification. In these pipelines, a first model predicts candidate values for CRF items, while a second model verifies whether the predictions are supported by the clinical note. Similarly, Greyc proposes a three-component architecture involving a rewriting step, an extraction stage, and a judging model that validates the predictions against the original note. These designs consistently aim to improve precision by filtering unsupported predictions.

Prompt engineering and task decomposition. Prompt design plays a key role in several submissions. SEBIS show through ablation studies that few-shot prompting improves performance and that

separating the prediction of item presence from value extraction further increases reliability. Innov8ors dynamically select in-context examples based on TF-IDF similarity to the input note, while Jovian_tech demonstrate the feasibility of predicting all CRF items simultaneously using a small set of fixed few-shot examples. Other teams explored task decomposition strategies. Aurum divide the CRF into domain-specific extractors implemented through DSPy signatures with chain-of-thought reasoning, enabling targeted extraction across clinical subdomains. Polimi instead design specialized prompts for different categories of items (binary, categorical, and measured variables), while Greyc employ k-shot prompting in an item-wise extraction framework. Overall, these approaches highlight how prompt design and task structuring can significantly influence extraction performance in CRF filling, and few-shot being an effective approach.

Hybrid LLM and rule-based systems. Several systems combine LLM predictions with rule-based post-processing. Polimi, for instance, augment prompts with a glossary of clinical abbreviations automatically extracted from unlabeled notes and apply pattern-based rules for items with identifiable textual cues. Gladiators similarly integrate rule-based components to address systematic extraction errors, including regular-expression extraction of vital signs, Unicode normalization for medical symbols, and explicit handling of symptom negation. DocUA adopts a comparable hybrid design, combining LLM-based extraction with deterministic post-processing for canonicalization, vocabulary normalization, and mapping to standardized medical concepts. These approaches highlight the continued value of domain knowledge and manual analysis of clinical text when designing reliable extraction pipelines.

8. Conclusion

In this paper, we presented an overview of the CRF filling Shared Task. A number of insights emerge across submissions. Overall, the task is far from being saturated, with the best performing system reaching a score of 68.3 and 67.1 for English and Italian respectively. Not surprisingly, systems powered by closed-source models perform better than open-source ones. Nevertheless, the gap can be significantly reduced by post-processing approaches based on manual or semi-automatic mappings. This highlights the importance of in-depth analysis of the dataset characteristics.

All participants faced the trade-off between precision and coverage, with systems that aggressively reduce false positives often sacrifice recall. To address this challenge, several teams report that man-

ual analysis of difficult items and targeted prompt or rule adjustments can yield measurable improvements, suggesting that future work may benefit from more fine-grained modeling of item-specific extraction patterns.

Overall, the shared task results indicate that while LLMs provide a strong baseline for CRF filling, robust performance often requires hybrid architectures combining prompting strategies, verification steps, and domain-specific post-processing.

9. Limitations

The main limitation of the Shared Task relies on the size of the gold-standard dataset. Overall, it comprises 290 clinical notes from a single Italian hospital. Future work should focus on increasing the size and diversifying the sources.

10. Acknowledgments

This work has been partially funded by the European Union under the Horizon Europe eCREAM Project (Grant Agreement No.101057726). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

11. Bibliographical References

- Shantala Bellary, Binny Krishnankutty, and M. S. Latha. 2014. [Basics of case report form designing in clinical research](#). *Perspectives in Clinical Research*, 5(4):159–166.
- Claudio Crema, Federico Verde, Pietro Tiraboschi, Camillo Marra, Andrea Arighi, Silvia Fostinelli, Guido Maria Giuffr , Vera Pacoova Dal Maschio, Federica L’Abbate, Federica Solca, Barbara Poletti, Vincenzo Silani, Emanuela Rotonondo, Vittoria Borracci, Roberto Vimercati, Valeria Crepaldi, Emanuela Inguscio, Massimo Filippi, Francesca Caso, Alessandra Maria Rosati, Davide Quaranta, Giuliano Binetti, Ilaria Pagnoni, Manuela Morreale, Francesca Burgio, Michelangelo Stanzani-Maserati, Sabina Capellari, Matteo Pardini, Nicola Girtler, Federica Piras, Fabrizio Piras, Stefania Lalli, Elena Perdixi, Gemma Lombardi, Sonia Di Tella, Alfredo Costa, Marco Capelli, Cira Fundar , Marina Manera, Cristina Muscio, Elisa Pellencin, Raffaele Lodi, Fabrizio Tagliavini, and Alberto Redolfi. 2024. [Medical information extraction with nlp-powered qabots: A real-world scenario](#). *IEEE Journal of Biomedical and Health Informatics*, 28(11):6906–6917.
- Pietro Ferrazzi, Mattia Franzin, Alberto Lavelli, and Bernardo Magnini. 2026. [Small LLMs for medical NLP: a systematic analysis of few-shot, constraint decoding, fine-tuning and continual pre-training in Italian](#).
- Pietro Ferrazzi, Alberto Lavelli, and Bernardo Magnini. 2025. [Converting annotated clinical cases into structured case report forms](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 307–318, Vienna, Austria. Association for Computational Linguistics.
- Alba Guti rrez-Sacrist n, Simran Makwana, Audrey Dionne, Simran Mahanta, Karla J. Dyer, Faridis Serrano, Carmen Watrin, Pierre Pages, Sajad Mousavi, Anil Degala, Jessica Lyons, Danielle Pillion, Joany M. Zachariasse, Lara S. Shekerdemian, Dongngan T. Truong, Jane W. Newburger, and Paul Avillach. 2024. [Development and validation of an open-source pipeline for automatic population of case report forms from electronic health records: a pediatric multi-center prospective study](#). *eBioMedicine*, 108. Doi: 10.1016/j.ebiom.2024.105337.
- Gabriela Anna Kaczmarek, Pietro Ferrazzi, Lorenzo Porta, Vicky Rubini, and Bernardo Magnini. 2026. [Toward automatic filling of case report forms: A case study on data from an italian emergency department](#).
- Ching-Heng Lin, Nai-Yuan Wu, and Der-Ming Liou. 2015. [A multi-technique approach to bridge electronic case report form design and data standard adoption](#). *Journal of Biomedical Informatics*, 53:49–57.
- W. R. Mac Kenzie, A. J. Davidson, A. Wiesenthal, J. P. Engel, K. Turner, L. Conn, S. J. Becker, S. Moffatt, S. L. Groseclose, J. Jellison, J. Stinn, N. Y. Garrett, L. Helmus, B. Harmon, C. L. Richards, J. R. Lumpkin, and M. F. Iademarco. 2016. [The promise of electronic case reporting](#). *Public Health Reports*, 131(6):742–746. Epub 2016 Oct 13.
- Maja Popovi . 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*.
- Maja Popovi . 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*.
- Frank P tavy et al. 2019. [Global standardization of clinical research data](#). *Applied Clinical Trials*, 28(4):20–23.

Reinhard Rapp. 2009. The backtranslation score: Automatic MT evaluation at the sentence level without reference translations. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*.

Rachel L. Richesson and Prakash Nadkarni. 2011. [Data standards for clinical research data collection forms: current status and challenges](#). *Journal of the American Medical Informatics Association*, 18(3):341–346.

Eugenia Rinaldi, Caroline Stellmach, and Sylvia Thun. 2025. [How to design electronic case report form \(ecrf\) questions to maximize semantic interoperability in clinical research](#). *Interactive Journal of Medical Research*, 14:e51598.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chu-jie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiayi Yang, 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](#). *CoRR*, abs/2505.09388.

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*.

Terry Yue Zhuo, Qionghai Xu, Xuanli He, and Trevor Cohn. 2023. Rethinking round-trip translation for machine translation evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*.

A. Appendix

A.1 Translation Prompt

Here is the system prompt utilized for automatic translation of the unannotated clinical notes:

```
You are a clinical-grade translation model specialized in emergency-department notes. Translate all text from Italian → English.
```

```
## Requirements
```

- Preserve all medical facts, nu-

```
meric values, units, abbreviations, and dates.
```

- Maintain the structure and conciseness typical of ED notes (headings, bullet points, line breaks).
- Use professional clinical English, not patient-friendly language.
- Do not summarize, interpret, add, infer, or remove information.
- Translate abbreviations to their standard English ED equivalents when unambiguous (e.g., PA → BP, FC → HR). If ambiguous, leave them unchanged.
- If a term is unclear, leave it unchanged.
- The notes have been anonymized: if you find terms such as NOME_PERSONA, LUOGO, PARENTE, NUM_TELEFONO leave them unchanged.

```
## Output
```

```
Return a json file with the English translation:
```

```
{
  "translation":
    "<full translated text here>"
}
```

This is the user prompt:

```
Translate the following ED note from Italian to English according to the system instructions: note
```

A.2 Baseline Prompt

Here we report the prompts used for the 2-shot Qwen3-8B baseline for both Italian and English.

System prompt:

```
You are a helpful medical assistant that extracts information in clinical notes based on patient history. Items must be filled with one of the valid values. If the information is not explicitly stated in the clinical note, answer 'unknown'.
```

User prompt:

Here is the patient history:

< < <{clinical_note}> > >

Fill this item: {item}

Valid values are: {valid_values}

Overview of the ArchEHR-QA 2026 Shared Task on Grounded Question Answering from Electronic Health Records

Sarvesh Soni, Dina Demner-Fushman

Division of Intramural Research

National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

sarvesh.soni@nih.gov, ddemner@mail.nih.gov

Abstract

We present an overview of the ArchEHR-QA 2026 Shared Task on grounded question answering from electronic health records (EHRs), organized at the CL4Health Workshop at LREC 2026. The 2026 task decomposes grounded EHR question answering (QA) into four complementary subtasks: question interpretation, evidence identification, answer generation, and evidence alignment. We evaluated submitted systems for the text-generation subtasks (question interpretation and answer generation) using lexical, semantic, and grounding-sensitive automatic metrics, and for the evidence-centric subtasks (evidence identification and evidence alignment) using precision, recall, and F1. The shared task received 198 submitted runs from 43 teams, and 17 teams additionally provided system descriptions for this overview. The highest-ranked systems differed across subtasks, and gains over the organizer baseline were largest on the evidence-centric subtasks. Across submitted system descriptions, prompt-based large language model (LLM) pipelines were dominant, whereas task-specific fine-tuning was rare; retrieval, self-consistency, and ensembling were especially common in the strongest evidence-centric systems. In this paper, we describe the task design, data, evaluation protocol, baselines, participation, official results, and common system characteristics, and discuss implications for developing clinically faithful and transparent QA systems.

Keywords: grounded question answering, electronic health records, clinical natural language processing, question reformulation, evidence attribution, shared task evaluation

1. Introduction

Patients frequently contact their healthcare providers with questions about diagnoses, treatments, test results, and events documented in their electronic health records (EHRs) (Tai-Seale et al., 2017). Patient messaging is also a substantial contributor to clinician burden (Yan et al., 2021; Martinez et al., 2024). From a clinical natural language processing (NLP) perspective, responding to a patient question from the EHR can be decomposed into four related steps: identifying the patient’s core information need, locating answer-bearing clinical evidence, drafting an answer, and linking the answer back to the supporting evidence. Systems that assist with any subset of these steps could reduce clinician burden while improving transparency.

Clinicians often find it challenging and time-consuming to process patient messages that are long, verbose, and emotionally charged (Lanham et al., 2018). Patients often ask longer questions, and as a result, prior work has studied summarization of consumer health questions that can be answered from general medical knowledge (Ben Abacha et al., 2021). However, comparatively less work has focused on reformulating patient-authored narratives into patient-specific clinical questions that can then be answered from the patient’s EHR (Soni and Demner-Fushman, 2025b).

EHR-grounded response drafting also requires

effective evidence identification. Because EHRs contain large amounts of heterogeneous patient information, finding the specific note sentences needed to answer a question is time-consuming (Koopman et al., 2015; Nijor et al., 2022). Although clinical information retrieval has been studied extensively, much of that literature concerns retrieval across patients or documents rather than extraction of patient-specific answer evidence within an individual patient’s EHR (Sivarajkumar et al., 2024).

Finally, clinicians prefer QA systems whose answers can be verified against clinical evidence (Kell et al., 2024). Grounding (linking text back to the source evidence that supports it) is therefore a central requirement for trustworthy clinical QA (Chandu et al., 2021). In the present setting, grounding should allow a clinician to inspect which parts of the EHR support each answer sentence, regardless of whether the draft answer was generated by a model, authored by a clinician, or jointly produced.

To study these interdependent capabilities, we organized the ArchEHR-QA 2026 Shared Task¹ as part of the CL4Health Workshop at LREC 2026. This is the second ArchEHR-QA shared task. The 2025 edition, organized at the BioNLP Workshop at ACL 2025, focused on end-to-end answer generation with citations to supporting note sentences (Soni et al., 2025). In contrast, the 2026 edition decomposes grounded EHR QA into four complemen-

¹archehr-qa.github.io

tary subtasks: question interpretation, evidence identification, answer generation, and evidence alignment. This decomposition permits a more fine-grained evaluation of the capabilities required for grounded response drafting. In this paper, we describe the task design, data, evaluation protocol, baseline, participation, official results, and observations from the submitted systems.

2. ArchEHR-QA 2026 Task Description

ArchEHR-QA 2026 consists of four subtasks: Question Interpretation, Evidence Identification, Answer Generation, and Evidence Alignment. Table 1 presents a running example illustrating the inputs and reference outputs for all four subtasks.

2.1. Subtask 1: Question Interpretation

Subtask 1 frames question interpretation as a constrained reformulation task. Given a free-text patient-authored question, the system must produce a concise clinician-interpreted question that captures the patient’s core clinical information need. The target output is a single well-formed question phrased as the type of query a clinician would submit to an intelligent EHR system when preparing a response. To preserve the focus on concise reformulation, outputs were limited to 15 words and were expected to avoid unsupported clinical additions. In the running example, the reference output is the clinician-interpreted question shown in Table 1. The clinician-interpreted question provides a concise reformulation of the patient’s core information need. For the downstream subtasks, however, we provided both the original patient-authored question and the clinician-interpreted question, since the two forms may contain complementary information, and participants could draw on either or both.

Given the patient question, the clinician-interpreted question, clinical specialty labels, and a sentence-segmented clinical note excerpt, the system must identify the note sentences that provide the evidence needed to answer the question.

2.2. Subtask 2: Evidence Identification

Subtask 2 frames evidence identification as sentence-level evidence selection. Given the patient question, the clinician-interpreted question, clinical specialty labels, and a sentence-segmented clinical note excerpt, the system must identify the note sentences that provide the evidence needed to answer the question. The target output is a set of sentence IDs corresponding to evidence that is sufficient for answering the question. This subtask emphasizes identification of essential evidence while

discouraging unnecessary or only tangentially related sentences. The reference evidence set for the running example is shown in Table 1.

2.3. Subtask 3: Answer Generation

Subtask 3 frames answer generation as grounded response drafting. Given the patient question, the clinician-interpreted question, clinical specialty labels, and the clinical note excerpt, the system must generate a concise answer supported by the note. Answers were limited to 75 words, roughly corresponding to a short paragraph of up to five sentences. We selected this limit based on preliminary baseline experiments and prior work suggesting that paragraph-length answers are generally preferred by users (Lin et al., 2003; Jeon et al., 2006). Reference answers were written in a professional clinical register and grounded in the note content, without unsupported speculation. Unlike Subtask 2, this subtask does not provide identified evidence sentences as input; instead, it evaluates answer generation directly from the note excerpt rather than from a pre-filtered evidence set. This keeps the task focused on grounded answer generation, which still requires locating relevant information within the note, while allowing teams to use evidence identified by their own systems, including predictions from Subtask 2. The reference answer for the running example is shown in Table 1.

2.4. Subtask 4: Evidence Alignment

Subtask 4 frames grounding as sentence-level answer-to-evidence alignment. Given the question context, clinical specialty labels, a sentence-numbered clinical note excerpt, and an answer to be grounded, the system must predict which note sentence(s) support each answer sentence. The target output is therefore a set of mappings from answer sentences to note sentence IDs. These alignments are many-to-many: one answer sentence may be supported by multiple note sentences, and one note sentence may support multiple answer sentences. Answer sentences without direct support may be assigned an empty evidence set. For the shared task, we intentionally used the clinician-authored reference answer as the input answer for this subtask. This design isolates the alignment problem from answer-generation errors and enables straightforward automatic evaluation against gold answer-to-evidence mappings. It also provides a controlled setting for developing methods that may later be extended to system-generated answers in end-to-end pipelines. In the running example, the gold output is the set of sentence-level answer-to-evidence mappings shown in Table 1.

ArchEHR-QA Patient Case
PQ Patient question: I just wrote about my dad given multiple shots of lasix after he was already so swelled his shin looked like it would burst open. Why would they give him so much. He was on oxygen and they took him off of the higher flow rate. CQ Clinician-interpreted question: Why was he given lasix and his oxygen flow rate was reduced? SP Clinical specialty label(s): Cardiology NOTE Clinical note excerpt: 1: Acute diastolic heart failure: Pt developed signs and symptoms of volume overload with shortness of breath, increased oxygen requirement and lower extremity edema. 2: Echo showed preserved EF, no WMA and worsening AI. 3: CHF most likely secondary to worsening valvular disease. 4: He was diuresed with lasix IV, intermittently on lasix gtt then transitioned to PO torsemide with improvement in symptoms, although remained on a small amount of supplemental oxygen for comfort. 5: Respiratory failure: The patient was intubated for lethargy and acidosis initially and was given 8 L on his presentation to help maintain his BP's. 6: This undoubtedly contributed to his continued hypoxemic respiratory failure. 7: He was advanced to pressure support with stable ventilation and oxygenation. 8: On transfer to the CCU patient was still intubated but off pressors. 9: Patient was extubated successfully. 10: He was reintubated transiently for 48 hours for urgent TEE and subsequently extubated without adverse effect or complication. REL Note Sentence Relevance Annotations: [1, 4, 5, 6] → <i>essential</i>; [3, 7] → <i>supplementary</i> ANS Reference answer (clinician-authored): A1: The patient was given Lasix for acute diastolic heart failure with symptoms including shortness of breath and lower extremity edema. A2: The patient was given 8 liters of fluid to help maintain his blood pressure which contributed to his respiratory failure. A3: After the patient's heart failure was treated with Lasix, he showed improvement in shortness of breath and his oxygen requirement, and he only needed to remain on a small amount of oxygen for comfort. ALN Answer to Note Sentence Alignments: A1 → [1, 4] A2 → [5, 6] A3 → [4]
ST1: Question interpretation
Inputs: PQ Patient question. Target output: CQ Clinician-interpreted question. <i>Sample system response (good):</i> Why were multiple doses of Lasix given despite decreased oxygen requirements? <i>Sample system response (bad – generic/malformed):</i> Lasix dosage protocol for severe edema with oxygen therapy?
ST2: Evidence identification
Inputs: PQ Patient question; CQ clinician-interpreted question; SP clinical specialty label(s); NOTE clinical note. Target output: REL Note Sentence Relevance Annotations <i>Sample system response (missing key evidence; P 1.00, R 0.50, F1 0.67):</i> [1, 4] <i>Sample system response (over-inclusive; P 0.50, R 1.00, F1 0.67):</i> [1, 4, 5, 6, 7, 8, 9, 10]
ST3: Answer generation
Inputs: PQ Patient question; CQ clinician-interpreted question; SP clinical specialty label(s); NOTE clinical note. Target output: ANS Reference answer. <i>Sample system response (bad – generic/not grounded):</i> Lasix is commonly used to remove excess fluid when patients are swollen or having trouble breathing, and oxygen can often be reduced once breathing improves. Changes in oxygen flow usually reflect better respiratory status. <i>Sample system response (ok – some grounding but mostly generic):</i> He had fluid overload from heart failure, so Lasix was used to reduce swelling and shortness of breath. It also reports improvement after diuresis, with only a small amount of supplemental oxygen needed for comfort.
ST4: Evidence alignment
Inputs: PQ Patient question; CQ clinician question; SP specialty; NOTE clinical note; ANS reference answer. Target output: ALN Answer to Note Sentence Alignments <i>Sample response (under-citing; P 0.75, R 0.60, F1 0.67):</i> A1: → [1]. A2: → [5]. A3: → [4, 5]. <i>Sample response (over-citing; P 0.50, R 1.00, F1 0.67):</i> A1: → [1, 4, 5]. A2: → [5, 6, 7]. A3: → [1, 4, 6, 7].

Table 1: Example showing an ArchEHR-QA patient case and the corresponding inputs and target outputs for all four subtasks, with representative sample system responses.

	Dev ($n=20$)	Test ($n=100$)	Test-2026 ($n=47$)
<i>Mean word count</i>			
Patient question	85.2	92.3	94.1
Clinician question	10.8	10.6	10.4
Note excerpt	320.8	380.4	497.2
Clinician answer	73.6	72.4	72.7
<i>Mean note sentences</i>			
Total	21.4	26.0	34.0
Essential	6.0 (28.3%)	6.6 (25.3%)	9.7 (28.6%)
Supplementary	1.3 (6.1%)	5.5 (21.3%)	6.9 (20.4%)
Not relevant	14.1 (65.7%)	13.9 (53.4%)	17.3 (51.0%)

Table 2: Dataset statistics by split. Values are mean word counts for text fields and mean sentence counts for note relevance categories; percentages denote proportions of note sentences within each split. Test-2026 is the official test split for Subtasks 1–3; Subtask 4 uses both Test and Test-2026.

3. Data Description

ArchEHR-QA 2026 builds on the ArchEHR-QA dataset introduced in prior work (Soni and Demner-Fushman, 2026). That work describes the dataset creation process, including the annotation schema and guidelines. The dataset contains patient cases drawn from intensive care and emergency department settings with hospitalization-related questions. Table 1 shows one illustrative example. The dataset pairs patient-authored questions, inspired by real patient information needs expressed in public health forums, with sentence-segmented clinical note excerpts derived from the MIMIC database. Each data instance is referred to as a patient case. Each case contains: (i) a free-text patient question; (ii) a clinician-interpreted question; (iii) a sentence-segmented clinical note excerpt; (iv) sentence-level relevance annotations for the note sentences (*essential*, *supplementary*, or *not relevant*); (v) a clinician-authored reference answer; (vi) sentence-level answer-to-evidence alignments; and (vii) one or more clinical specialty labels.

ArchEHR-QA 2026 uses 167 patient cases. Of these, 134 cases come from the original release (Soni and Demner-Fushman, 2026), and 33 cases were newly curated for the 2026 shared task using the same annotation schema.

3.1. Official Evaluation Splits and Staged Release Schedule

The official test set for Subtasks 1–3 comprised case IDs 121–167 (47 cases), while case IDs 1–120 were released for system development. We restricted official evaluation of Subtasks 1–3 to case IDs 121–167 because case IDs 1–120 were already publicly used in the 2025 iteration of ArchEHR-QA (Soni et al., 2025). For Subtask 1, the prior release of clinician-interpreted questions made those cases unsuitable for a new evaluation. We used the same

held-out split for the closely related Subtasks 2–3 to avoid reusing cases that had already been used in a similar shared task setting in 2025. By contrast, Subtask 4 was newly introduced in 2026 and had not previously been evaluated on those cases. We therefore evaluated Subtask 4 on the larger split of case IDs 21–167 (147 cases), reserving case IDs 1–20 for development. Table 2 summarizes dataset statistics across the splits.

The official evaluation followed a staged release schedule designed to mirror the intended workflow and preserve separation between subtasks. For Subtask 1, participants initially received only the patient-authored questions for the 47 held-out test cases. After the Subtask 1 deadline, the release for Subtasks 2–3 added the clinician-interpreted questions and clinical note excerpts for the same 47 cases. After the Subtasks 2–3 deadline, the release for Subtask 4 provided case IDs 21–167 together with sentence-numbered reference answers to be aligned to the corresponding note excerpts. Sentence-level relevance annotations were withheld from all official test releases.

4. Evaluation

4.1. Metrics

Each subtask was evaluated independently on its corresponding held-out test set. Official leaderboard rankings were determined by a single primary metric per subtask, while additional metrics were reported to provide a broader characterization of system behavior. The scoring scripts are available on GitHub².

Subtask 1. The system-generated clinician-interpreted questions were compared with the reference questions using ROUGE-Lsum (Lin, 2004),

²github.com/soni-sarvesh/archehr-qa-2026

BERTScore (Zhang et al., 2019), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023). These metrics capture complementary aspects of reformulation quality: ROUGE-Lsum measures lexical overlap, BERTScore measures semantic similarity, AlignScore measures factual consistency, and MEDCON measures clinical concept overlap. To enforce the task constraint, outputs longer than 15 words were truncated to the first 15 words before scoring. The official ranking metric was the Overall Score (OS), computed as the unweighted arithmetic mean of the four metric values.

Subtask 2. The predicted evidence sentence IDs were compared with the reference evidence annotations using Precision, Recall, and F1. We report both micro-averaged scores, which pool predictions across all cases, and macro-averaged scores, which average case-level scores. Results are reported under two evaluation variants. Under *Strict* evaluation, only sentences labeled *essential* are treated as gold evidence. Under *Lenient* evaluation, predicted *supplementary* sentences were not counted as false positives. The official ranking metric was Strict Micro F1.

Subtask 3. The system-generated answers were compared with the reference answers using BLEU (Papineni et al., 2002), ROUGE-Lsum, SARI (Xu et al., 2016), BERTScore, AlignScore, and MEDCON. Together, these metrics capture complementary aspects of answer quality, with BLEU and ROUGE-Lsum emphasizing lexical overlap, SARI capturing content-editing behavior, and BERTScore, AlignScore, and MEDCON providing semantic, factual, and clinical-concept perspectives. Prior work in this setting suggests that automatic metrics can provide useful system-ranking signals when computed against clinician-authored reference answers (Soni and Demner-Fushman, 2025a). The official ranking metric was the Overall Score (OS), computed as the unweighted arithmetic mean of the six metric values.

Subtask 4. The predicted alignments were evaluated as sets of links between answer sentences and note sentences, where each link has the form (*answer sentence k* $\rightarrow$ *note sentence i*). We report Precision, Recall, and F1 at both the micro and macro levels. Micro-averaged scores pool predicted and reference links across all cases, whereas macro-averaged scores average case-level scores. Over-citing is penalized through false positives, which lowers Precision and therefore F1. The official ranking metric was Micro F1.

4.2. Baseline

We used a zero-shot baseline based on Qwen/Qwen3-4B-Instruct-2507 (Yang et al., 2025) for all four subtasks. No task-specific fine-tuning or few-shot exemplars were used. The model was prompted in a single-turn chat format with task-specific instructions; the prompt templates are provided in Appendix Tables 8–11. We used the same decoding settings across subtasks, following the Qwen3-recommended hyperparameters³: temperature = 0.7, top-*p* = 0.8, top-*k* = 20, and min-*p* = 0.0.

For Subtask 1, the model rewrote each patient-authored question as a clinician-interpreted question (`max_new_tokens=256`). For Subtask 2, the model predicted supporting sentence IDs from the clinical note excerpt conditioned on the patient-authored question, the clinician-interpreted question, and the specialty label (`max_new_tokens=512`). For Subtask 3, the model generated a clinically grounded answer from the patient-authored question, the clinician-interpreted question, the specialty label, and the clinical note excerpt (`max_new_tokens=512`). For Subtask 4, the model predicted answer-to-evidence sentence alignments from the clinician-interpreted question, the sentence-numbered clinical note excerpt, and the sentence-numbered answer (`max_new_tokens=1024`). Although both question forms were available for Subtask 4, the baseline used only the clinician-interpreted question, based on preliminary experiments on the development set.

5. Participation

We used Codabench⁴ to collect system predictions from teams (Xu et al., 2022). Teams could participate in any subset of the four subtasks and were allowed to submit up to three runs per subtask. For the overview paper, we asked each team to nominate its best run per subtask and to submit a short system description.

5.1. Participating Teams

We received 198 submitted runs from 43 teams across the four subtasks: 41 runs for Subtask 1, 59 for Subtask 2, 35 for Subtask 3, and 63 for Subtask 4. Among these 43 teams, 17 submitted system descriptions and are therefore included in the

³huggingface.co/Qwen/Qwen3-4B-Instruct-2507

⁴Subtask 1: codabench.org/competitions/12865;
Subtask 2: codabench.org/competitions/13526;
Subtask 3: codabench.org/competitions/13527;
Subtask 4: codabench.org/competitions/13528

Team ID	Affiliation	Country	ST1	ST2	ST3	ST4
BIT.UA-AAUBS	University of Aveiro; Aalborg University	Portugal; Denmark	✓	✓	✓	✓
CaresAI	Cairo University; Techvify	Egypt; Vietnam	✓	✓	×	×
GigitAI	GigitAI	USA	×	✓	✓	×
HealthNLP_Retrievers	University of Maryland Baltimore County; Jahangirnagar University	USA; Bangladesh	✓	✓	✓	✓
HiTZ-IXA	University of the Basque Country	Spain	×	✓	×	✓
KA	L&T Technology Services	India	×	×	×	✓
KPSCMI	Kaiser Permanente Southern California	USA	✓	✓	×	✓
MedEvi-NS	University of Aberdeen; Bournemouth University	UK	×	×	×	✓
Neural	University of Chicago; University of Wisconsin Madison; Lovely Professional University	USA; India	✓	✓	✓	✓
OptiMed	Queen Mary University of London; Independent Researcher	UK; Turkey	✓	✓	✓	✓
razreshili	Independent Researcher	Germany	×	×	×	✓
sebis	Technical University of Munich	Germany	✓	✓	✓	✓
TAMU-NLP-Lab	Texas A&M University	USA	×	✓	✓	×
tt501	Vietnam National University	Vietnam	×	✓	✓	✓
UIC-AIHealth4All	University of Illinois Chicago	USA	×	✓	✓	✓
WisPerMed	University of Applied Sciences and Arts Dortmund; University of Duisburg-Essen	Germany	✓	✓	✓	✓
Yale-DM-Lab	Yale University; UMass Lowell	USA	✓	✓	✓	✓

Table 3: Participating teams who submitted a system description, their affiliation, country, and the subtasks they participated in. Rows are sorted by Team ID in alphabetical order. ST1–ST4 denote Subtasks 1–4.

analysis below. Table 3 lists those teams, their affiliations, countries, and the subtasks they entered. Among the teams represented in the overview paper, seven participated in all four subtasks.

5.2. Results

Tables 4–7 report the official results for subtasks 1–4. All highest-ranked submissions outperformed the organizer baseline, but the margin varied substantially across subtasks: +11.3 Overall Score for Subtask 1 (31.2 vs. 19.9), +14.1 Strict Micro F1 for Subtask 2 (63.7 vs. 49.6), +4.3 Overall Score for Subtask 3 (36.3 vs. 32.0), and +16.2 Micro F1 for Subtask 4 (81.5 vs. 65.3). The baseline remained competitive with lower-ranked submissions in Subtasks 1 and 3, whereas the evidence-centric subtasks showed a clearer separation between the strongest submitted systems and the zero-shot baseline.

For Subtask 1 (Question Interpretation; Table 4), the highest-ranked system was submitted by *HealthNLP_Retrievers* and obtained an overall score of 31.2, followed by *KPSCMI* (30.8) and *OptiMed* (29.9). The top three systems were

★ Team	Relevance				
	RG	BS	AS	MD	OS
1 HealthNLP_Retrievers	35.3	46.8	24.0	18.7	31.2
2 KPSCMI	27.8	41.0	26.4	27.9	30.8
3 OptiMed	28.8	43.1	27.7	19.9	29.9
4 Neural	31.3	43.6	15.2	25.6	28.9
5 Yale-DM-Lab	28.2	40.6	19.7	19.8	27.1
6 WisPerMed	21.8	38.0	21.7	26.3	26.9
7 sebis	22.9	36.9	21.4	21.3	25.6
8 CaresAI	22.5	36.1	12.5	24.5	23.9
9 BIT.UA-AAUBS	17.9	29.4	8.7	20.2	19.0
- baseline	23.4	31.0	7.9	17.3	19.9

Table 4: Subtask 1 (Question Interpretation) results. Rows are ranked (★) by the overall score (OS). RG: ROUGE-Lsum; BS: BERTScore; AS: AlignScore; MD: MEDCON; OS: Overall score.

separated by only 1.3 overall-score points. The strongest systems emphasized different aspects of the task: the *HealthNLP_Retrievers* submission obtained the highest ROUGE-Lsum and BERTScore,

Rank	Team	Micro						Macro					
		Strict			Lenient			Strict			Lenient		
		P	R	F1 [†]	P	R	F1	P	R	F1	P	R	F1
1	Neural	60.2	67.6	63.7	77.8	67.6	72.4	64.3	69.7	64.8	81.9	69.7	73.1
2	OptiMed	56.7	71.3	63.2	73.1	71.3	72.2	61.5	73.0	64.1	78.1	73.0	72.9
3	UIC-AIHealth4All	59.3	67.0	62.9	74.3	67.0	70.4	61.8	69.2	63.0	77.8	69.2	70.5
4	Yale-DM-Lab	52.3	75.7	61.9	68.0	75.7	71.6	56.7	75.5	61.1	73.8	75.5	70.4
5	TAMU-NLP-Lab	56.7	65.9	60.9	76.0	65.9	70.6	63.0	66.7	59.9	80.0	66.7	69.1
6	HealthNLP_Retrievers	58.4	62.1	60.2	75.7	62.1	68.3	65.7	62.5	59.9	79.8	62.5	67.1
7	KPSCMI	46.5	81.8	59.3	62.4	81.8	70.8	51.9	82.2	60.9	69.4	82.2	72.0
8	tt501	56.0	61.9	58.8	74.9	61.9	67.8	61.3	63.9	58.4	78.4	63.9	66.8
9	WisPerMed	67.1	52.3	58.8	84.2	52.3	64.5	69.8	53.4	58.1	84.9	53.4	63.6
10	BIT.UA-AAUBS	51.0	69.4	58.8	68.5	69.4	68.9	59.8	69.1	59.3	76.9	69.1	68.9
11	GigitaI	46.5	77.0	58.0	62.1	77.0	68.8	52.8	77.7	59.1	69.6	77.7	69.3
12	CaresAI	46.7	67.6	55.3	61.3	67.6	64.3	50.1	68.2	55.2	66.8	68.2	64.0
13	sebis	37.9	80.7	51.6	49.1	80.7	61.1	40.8	80.3	52.6	54.1	80.3	62.6
14	HiTZ-IXA	44.2	47.9	46.0	56.3	47.9	51.8	48.7	48.7	46.7	62.4	48.7	52.2
-	baseline	51.0	48.1	49.6	64.5	48.1	55.1	60.7	47.2	47.9	74.8	47.2	52.8

Table 5: Subtask 2 (Evidence Identification) results. Rows are ranked by Strict Micro F1 Score ([†]). P: Precision; R: Recall; F1: F1 score.

Rank	Team	Lexical			Semantic			Overall
		BL	RG	SA	BS	AS	MD	OS
1	WisPerMed	9.9	27.8	58.6	46.8	31.7	43.1	36.3
2	TAMU-NLP-Lab	9.7	27.4	59.6	46.2	34.5	39.9	36.2
3	BIT.UA-AAUBS	8.6	26.4	60.0	45.0	30.2	43.2	35.6
4	Neural	9.4	25.6	57.7	43.5	34.3	40.9	35.2
5	HealthNLP_Retrievers	7.0	25.4	59.2	43.8	33.6	38.7	34.6
6	OptiMed	5.7	25.2	56.5	43.1	37.4	39.1	34.5
7	UIC-AIHealth4All	6.1	24.2	57.2	41.8	24.5	37.5	31.9
8	GigitaI	5.1	21.1	56.8	40.0	28.4	39.4	31.8
9	sebis	3.8	22.0	56.6	39.3	33.9	33.4	31.5
10	tt501	6.2	24.0	58.3	41.4	22.4	35.9	31.4
11	Yale-DM-Lab	9.1	23.1	56.5	37.2	22.1	37.6	31.0
-	baseline	5.0	22.5	56.2	40.6	32.2	35.4	32.0

Table 6: Subtask 3 (Answer Generation) results. Rows are ranked by the overall score (OS). BL: BLEU; RG: ROUGE-Lsum; SA: SARI; BS: BERTScore; AS: AlignScore; MD: MEDCON; OS: Overall score.

the *OptiMed* submission obtained the highest AlignScore, and the *KPSCMI* submission obtained the highest MEDCON.

In Subtask 2 (Evidence Identification; Table 5), the *Neural* system achieved the best Strict Micro F1 (63.7), narrowly ahead of *OptiMed* (63.2) and *UIC-AIHealth4All* (62.9). Lenient evaluation improved F1 for all systems, indicating that many submissions selected supplementary but still clinically plausible evidence sentences. The systems showed different precision–recall profiles: *KPSCMI* and *sebis* favored higher recall, whereas *Neural* and *OptiMed* achieved a more balanced trade-off.

For Subtask 3 (Answer Generation; Table 6), the *WisPerMed* system obtained the highest overall score (36.3), with *TAMU-NLP-Lab* effectively tied at 36.2 and *BIT.UA-AAUBS* third at 35.6. Metric leaders again differed by evaluation perspective: the *WisPerMed* submission obtained the best BLEU, ROUGE-Lsum, and BERTScore; the *BIT.UA-AAUBS* submission obtained the best SARI and MEDCON; and the *OptiMed* submission obtained the best AlignScore. This result suggests that answer generation systems are ranked differently depending on whether the evaluation emphasizes lexical overlap, semantic similarity, or clinical

Rank	Team	Micro			Macro		
		P	R	F1 [†]	P	R	F1
1	BIT.UA-AAUBS	88.0	75.9	81.5	88.3	78.9	82.2
2	WisPerMed	86.9	76.3	81.3	87.5	79.4	82.1
3	Yale-DM-Lab	83.3	77.7	80.4	84.8	80.7	81.8
4	OptiMed	80.7	79.8	80.3	82.8	82.8	81.6
5	UIC-AIHealth4All	83.6	76.3	79.8	85.5	79.5	80.9
6	tt501	78.2	80.1	79.1	80.0	82.6	79.7
7	Neural	84.3	73.7	78.6	85.8	76.6	79.4
8	KPSCMI	86.2	71.5	78.1	86.4	75.5	78.9
9	HealthNLP_Retrievers	83.8	71.1	76.9	85.3	74.1	77.4
10	MedEvi-NS	75.7	77.4	76.5	79.3	80.8	78.3
11	sebis	79.6	70.6	74.8	82.6	73.6	76.2
12	razreshili	73.8	63.0	67.9	75.8	67.6	69.5
13	HiTZ-IXA	74.7	60.3	66.8	73.5	65.4	66.7
14	KA	68.1	64.5	66.2	70.9	69.0	68.0
-	baseline	59.1	72.9	65.3	72.0	75.3	71.1

Table 7: Subtask 4 (Evidence Alignment) results. Rows are ranked by Micro F1 Score ([†]). P: Precision; R: Recall; F1: F1 score.

concept preservation. Compared with the other subtasks, Subtask 3 showed a relatively small margin over the organizer baseline (36.3 vs. 32.0).

For Subtask 4 (Evidence Alignment; Table 7), the highest-ranked system was submitted by *BIT.UA-AAUBS* and obtained a Micro F1 of 81.5, followed closely by *WisPerMed* (81.3), *Yale-DM-Lab* (80.4), and *OptiMed* (80.3). The top four systems were separated by only 1.2 F1 points. Precision was generally high among the strongest submissions, while Recall accounted for most of the remaining variation.

Among teams participating in all four subtasks, *OptiMed* showed the most consistently strong performance, placing within the top six in every subtask. No single team submitted the highest-ranked system in more than one subtask.

5.3. Approaches

Across all four subtasks, prompt-based use of large language models (LLMs) was dominant. Task-specific fine-tuning was comparatively rare; more common strategies were prompt engineering, retrieval of similar development examples, repeated sampling with self-consistency, and multi-model ensembling. Strong submissions were obtained with both API-based models and locally hosted or open-weight models, suggesting that no single deployment setup dominated the task.

For Subtask 1 (Question Interpretation), in-context prompting with examples was frequent: 8/9 systems used few-shot or many-shot prompting, often with explicit post-processing to enforce the

single-question output format and the 15-word limit. Five of the nine systems used proprietary API-based LLMs, whereas the remainder relied on locally hosted models. Three systems introduced an additional reasoning or decomposition step before generation, e.g., by extracting important medical terms or by scoring candidate questions from multiple models. Prompt optimization was reported by two systems, using frameworks such as MIPROv2 (Opsahl-Ong et al., 2024) or GEPA (Agrawal et al., 2025). The strongest systems on this subtask followed three different designs: few-shot prompting with tight output constraints on an API-based model (*HealthNLP_Retrievers*), many-shot prompting with the full development set as examples on a locally hosted model (*KPSCMI*), and zero-shot generation via clinical-situation abstraction and prompt optimization (*OptiMed*).

For Subtask 2 (Evidence Identification), most systems used both the patient-authored and clinician-interpreted questions as input (11/14). Zero-shot prompting was marginally more common than few-shot prompting (7/14 vs. 6/14), and API-based models dominated (10/14). Methodologically, most submissions framed evidence identification either as direct extraction of supporting sentence IDs from the note excerpt (10/14) or as sentence-wise classification (3/14). Verification-based strategies were common: five systems used self-reflection or iterative refinement, two used self-consistency across multiple generations, and four used multi-model ensembling. Three systems also reported prompt optimization, while one system departed from the dominant prompting-based methods by

using embedding-based sentence similarity. A recurring theme was recall optimization: several systems used inclusive prompts, answer-first reasoning, union voting, or post-hoc refinement to reduce the risk of omitting essential evidence. The highest-ranked systems illustrate two contrasting but effective strategies: calibrated sentence-level classification with verification and self-consistency (*Neural*), and zero-shot multi-model ensembling with majority voting (*OptiMed*).

For Subtask 3 (Answer Generation), both few-shot and zero-shot prompting were common, with few-shot setups only marginally more common than zero-shot setups (6/11 vs. 5/11). API-based models dominated this subtask (9/11), and most systems used both the patient-authored and clinician-interpreted questions at inference time (9/11). A common design choice was to condition generation on a filtered evidence subset, often obtained from the team's Subtask 2 pipeline, rather than on the full note excerpt alone (6/11 systems). Two systems used model ensembles, three used self-reflection or iterative self-revision, and three optimized prompts on the development set. Several teams adopted multi-stage generation procedures, including candidate generation followed by LLM-based selection, generate–critique–revise pipelines, and cite–then–rewrite strategies. As in the other subtasks, explicit task-specific fine-tuning was rare. The strongest reported systems again reflect different design choices: dynamic example retrieval with few-shot prompting (*WisPerMed*), intent-consistent in-context example selection (*TAMU-NLP-Lab*), and evidence-grounded multi-model generation with LLM-based selection (*BIT.UA-AAUBS*).

For Subtask 4 (Evidence Alignment), few-shot prompting was somewhat more common than zero-shot prompting (8/14 vs. 6/14). Nine of the fifteen systems used API-based models, while the remainder relied exclusively on locally hosted models. Input configurations were more varied than in Subtask 3: 7/14 systems used both question forms, 4 used only the patient question, and 3 used only the clinician-interpreted question. There was greater emphasis on precision–recall calibration through voting thresholds, self-consistency, and prompt instructions that explicitly favored either more conservative or more inclusive linking. Three systems used majority-vote ensembling across models, five used self-consistency, and only two explicitly reported prompt optimization. A small number of systems supplemented LLM predictions with embedding-based heuristics or fine-tuned re-ranking components. The highest-ranked systems show that strong performance can be achieved with either few-shot majority-vote ensembling (*BIT.UA-AAUBS*), zero-shot prompt-only alignment (*Wis-*

PerMed), or ensemble prediction augmented with embedding-based recall heuristics (*Yale-DM-Lab*).

Taken together, the submitted system descriptions suggest that the benchmark decomposition influenced not only evaluation but also system design. Several teams explicitly propagated information across stages, e.g., by feeding predicted evidence into answer generation or by separating evidence identification from answer generation and alignment.

6. Conclusion

We presented an overview of the ArchEHR-QA 2026 Shared Task on grounded question answering from electronic health records. By decomposing the problem into question interpretation, evidence identification, answer generation, and evidence alignment, the benchmark enables separate evaluation of the main capabilities required for grounded response drafting to patient questions. The shared task attracted 198 submitted runs from 43 teams, with 17 teams contributing system descriptions for this overview. The highest-ranked systems differed across subtasks, and the largest gains over the organizer baseline were observed for evidence identification and evidence alignment. Across submitted systems, prompt-based LLM pipelines were dominant, task-specific fine-tuning was rare, and the strongest evidence-centric systems commonly combined retrieval, self-consistency, iterative refinement, and ensembling. We expect ArchEHR-QA 2026 to support future work on clinically faithful and verifiable QA systems for patient–clinician communication.

7. Limitations

This study has several limitations. First, the official test sets are modest in size, especially for Subtasks 1–3 (47 cases), so small numerical differences between closely ranked systems should not be over-interpreted. Second, the qualitative method analysis is limited to teams that submitted system descriptions; the remaining participating teams are reflected in the overall participation counts but not in the approach analysis. Third, evaluation relies primarily on automatic metrics. Although these metrics provide broad coverage of lexical and semantic similarity, they do not fully capture clinical usefulness, communication quality, or patient safety. This issue is especially relevant for Subtasks 1 and 3, where a single reference output may not capture the full range of clinically acceptable responses, and reference-based metrics may therefore under-reward valid paraphrases. Finally, the benchmark evaluates subtasks independently and uses curated note excerpts rather than the full

longitudinal EHR. Consequently, it does not fully capture end-to-end error propagation, open-ended retrieval over complete patient records, or generalization to other care settings such as outpatient longitudinal messaging.

8. Acknowledgements

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. We thank all participating teams for their submissions and for contributing system descriptions used in this overview.

9. Bibliographical References

- Lakshya A. Agrawal, Shangyin Tan, Dilara Soyulu, Noah Ziemis, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2025. GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning. In *First Workshop on Foundations of Reasoning in Language Models*.
- 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.
- Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. [Grounding ‘Grounding’ in NLP](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.
- Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, and Soyeon Park. 2006. [A framework to predict the quality of answers with non-textual features](#). In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’06, pages 228–235, New York, NY, USA. Association for Computing Machinery.
- Gregory Kell, Angus Roberts, Serge Umansky, Linglong Qian, Davide Ferrari, Frank Soboczenski, Byron C Wallace, Nikhil Patel, and Iain J Marshall. 2024. [Question answering systems for health professionals at the point of care—a systematic review](#). *Journal of the American Medical Informatics Association*, 31(4):1009–1024.
- Richelle J. Koopman, Linsey M. Barker Steege, Joi L. Moore, Martina A. Clarke, Shannon M. Canfield, Min S. Kim, and Jeffery L. Belden. 2015. [Physician Information Needs and Electronic Health Records \(EHRs\): Time to Reengineer the Clinic Note](#). *The Journal of the American Board of Family Medicine*, 28(3):316–323.
- Holly Jordan Lanham, Luci K. Leykum, and Jacqueline A. Pugh. 2018. [Examining the Complexity of Patient-Outpatient Care Team Secure Message Communication: Qualitative Analysis](#). *Journal of Medical Internet Research*, 20(7):e9269.
- 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.
- Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, and David R Karger. 2003. What Makes a Good Answer? The Role of Context in Question Answering. In *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction*, Zurich, Switzerland.
- Kathryn A. Martinez, Rebecca Schulte, Michael B. Rothberg, Maria Charmaine Tang, and Elizabeth R. Pfoh. 2024. [Patient Portal Message Volume and Time Spent on the EHR: An Observational Study of Primary Care Clinicians](#). *Journal of General Internal Medicine*, 39(4):566–572.
- Sohn Nijor, Gavin Rallis, Nimit Lad, and Eric Gokcen. 2022. [Patient Safety Issues From Information Overload in Electronic Medical Records](#). *Journal of Patient Safety*, 18(6):e999.
- Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. [Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366, Miami, Florida, USA. 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 on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.
- 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. 2025a. [Automated Evaluation can Distinguish the Good and Bad AI Responses to Patient Questions about Hospitalization](#).
- Sarvesh Soni and Dina Demner-Fushman. 2025b. Automatically Generating Patient-specific Clinical Questions from Patient Messages to Relieve Clinician Burden. In *AMIA Annual Symposium Proceedings*, Atlanta, Georgia. American Medical Informatics Association.
- Sarvesh Soni and Dina Demner-Fushman. 2026. [A Dataset for Addressing Patient’s Information Needs related to Clinical Course of Hospitalization](#). *Scientific Data*.
- 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.
- 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.
- 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.
- 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.
- 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.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chu-jie 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, Jiayi 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](#).
- 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 Zhit-ing 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. Appendix

You are an expert clinical natural language processing (NLP) assistant. Your task is to transform a patient-authored question into a clear and concise clinician-interpreted question.

You are given:

- a patient-authored question.

Your goal:

- Write a clinician-interpreted question that captures the core clinical information need implied by the patient-authored question.
- Phrase it as a query a clinician would ask a smart electronic health record (EHR) system to retrieve relevant chart information needed to answer the patient.
- Keep it concise (maximum 15 words).
- Preserve the patient's core concern without introducing new clinical facts not explicitly stated in the patient's narrative.

Output only the clinician-interpreted question, nothing else.

Patient question:

{__}

Table 8: Zero-shot prompt used for the Subtask 1 baseline.

You are an expert clinical natural language processing (NLP) assistant. Your task is to select the sentence(s) in a clinical note excerpt that contain the clinical evidence needed to answer a patient's question.

You are given:

- A patient-authored question.
- A clinician-interpreted version of the patient question.
- The clinical specialty(ies) relevant to the patient's question (may differ from the note's specialty; comma-separated).
- A clinical note excerpt with numbered sentences.

Your goal:

- Select sentence IDs from the note excerpt that provide sufficient clinical evidence to answer the patient's question.
- Include all sentences necessary to fully support the answer (e.g., event + date; medication name + dose).
- Only include sentences that are directly relevant and answer-bearing, including necessary qualifiers (e.g., dates, values, negations, attribution).
- Do not include sentences that are unrelated, generic, purely administrative, or only tangentially related.
- If there is no sentence in the excerpt that provides evidence to answer the question, output an empty array: [].
- You may select as many sentences as needed to provide sufficient evidence; the entire excerpt may not be required.

Output only a JSON array of sentence ID numbers (as integers), nothing else. For example: [1, 5, 6]

Patient question:

{__}

Clinician-interpreted question:

{__}

Clinical specialty(ies) of the question (comma-separated):

{__}

Clinical note excerpt (with numbered sentences):

{__}

Table 9: Zero-shot prompt used for the Subtask 2 baseline.

You are an expert clinical natural language processing (NLP) assistant. Your task is to generate a concise, clinically grounded answer to a patient's question using only the information in the provided clinical note excerpt.

You are given:

- A patient-authored question.
- A clinician-interpreted version of the patient question.
- The clinical specialty(ies) relevant to the patient's question (may differ from the note's specialty; comma-separated).
- A clinical note excerpt.

Your goal:

- Answer the patient's question using only information found in the clinical note excerpt.
- Write in a professional clinical register (not simplified lay language).
- Be concise: limit your answer to 75 words (approximately 5 sentences).
- Stay faithful to the clinical note -- do not speculate or add information not supported by the note.
- If the note does not contain sufficient information to fully answer the question, provide a faithful response based on what is available without speculation.
- Do not use outside medical knowledge or inference. Only use what is explicitly stated in the note excerpt.

Output only the answer text, nothing else.

Patient question:

{__}

Clinician-interpreted question:

{__}

Clinical specialty(ies) of the question (comma-separated):

{__}

Clinical note excerpt:

{__}

Table 10: Zero-shot prompt used for the Subtask 3 baseline.

You are an expert clinical natural language processing (NLP) assistant. Your task is to align each answer sentence to the clinical note sentence(s) that support it.

You are given:

- A clinician-interpreted version of the patient question.
- A clinical note excerpt with numbered sentences.
- An answer with numbered sentences.

Your goal:

- For each answer sentence, select the clinical note sentence(s) that directly support it.
- If an answer sentence is supported by multiple note sentences, include all supporting note sentences. Assign an empty list only if that answer sentence is not supported by any note sentence.
- Each answer sentence may be supported by zero, one, or multiple note sentences.

Output a JSON array where each element has "answer_id" (string) and "evidence_id" (array of strings). Output nothing else.

Example output format:

```
[{"answer_id": "1", "evidence_id": ["2"]}, {"answer_id": "2", "evidence_id": ["5", "6"]}, {"answer_id": "3", "evidence_id": []}]
```

Clinician-interpreted question:

{__}

Clinical note excerpt (numbered sentences):

{__}

Answer (numbered sentences, no citations):

{__}

Table 11: Zero-shot prompt used for the Subtask 4 baseline.

Structured Radiology Intelligence: Extracting Structured Data from MRI Reports Using LLMs

**Sushvin Marimuthu, Parameswari Krishnamurthy, Dipti Misra Sharma
Goldwin H, Anu Eapen, Betty Simon, Anuradha Chandramohan**

IIIT Hyderabad, IIIT Hyderabad, IIIT Hyderabad,
CMC Vellore, CMC Vellore, CMC Vellore, CMC Vellore
sushvin.m@research.iiit.ac.in, param.krishna@iiit.ac.in, dipti@iiit.ac.in,
change2.dm@gmail.com, anuepn@yahoo.com, drbettysimon@gmail.com,
Anuradha.Chandramohan@cmcvellore.ac.in

Abstract

This study presents efforts focused on extracting and structuring doctor notes, specifically Magnetic Resonance Imaging (MRI) reports, into a standardized format using large language models (LLMs). We introduce a novel benchmark dataset comprising of 55 clinically relevant variables given by doctors, making it the first of its kind in the automated processing of unstructured medical texts. The annotations to the dataset were generated using a systematic prompt-tuning approach that was manually validated. It was then evaluated across three experimental stages: baseline, intermediate, and fine-tuned. Each stage assessed the impact of different prompt strategies on the performance of various LLMs (LLaMA, Qwen, and DeepSeek). Among the models tested, LLaMA 3.1 8B Instruct consistently achieved the highest composite score in both the intermediate and final phases, resulting in an 18.42% improvement in performance.

Keywords: Unstructured to Structured, MRI Report, LLM, schema-guided extraction

1. Introduction

In healthcare, a Magnetic Resonance Imaging (MRI) radiology report is a detailed medical document written by a radiologist after analyzing MRI scan images (Hashemi et al., 2012). It describes the internal structures of the body, highlights any abnormalities, and provides a professional interpretation to guide further medical care. These reports are essential because they convert complex MRI images into meaningful insights that help the referring doctors to make accurate diagnoses, plan treatments, or monitor recovery. The report typically includes sections such as clinical history, findings, and impression, and it plays a key role in the patient's overall medical management (John and Mittal, 2025). While most MRI reports follow a general format, the doctor notes based on these reports are usually written in a semi-structured or unstructured manner, meaning the content is expressed in free-form language rather than fixed templates.

The objective of this study is to develop a data extraction system¹ that converts such semi-structured or unstructured doctor notes on MRI reports into a structured and standardized format, such as JSON. The system will identify and extract key clinical elements from the notes and organize them in a consistent structure. This structured data can then be used for various downstream applications, including integration with hospital information systems, support for clinical decision-making, large-scale

data analysis, and the generation of tabular or other standardized report formats.

To develop this system, we began by constructing a novel benchmark dataset consisting of 55 clinically relevant variables defined by medical professionals. The details of the base dataset, along with its annotation process and validation using multiple LLMs under various settings, is described in detail in Section 2.

1.1. Related Works

Multiple efforts in the recent have focused on converting semi-structured and unstructured medical texts into structured data formats. For instance, (Ntinopoulos et al., 2025) investigates the use of large language models (LLMs) to automate entity extraction and classification from health records, enabling faster generation of structured data from clinical narratives. Similarly, (Ying et al., 2025) introduced GENIE, a Generative Note Information Extraction System that leverages LLMs to transform health records into standardized data formats. These two studies differ notably in the number of variables extracted and the LLMs employed. Additionally, (Skyles et al., 2025) explored structured data extraction from oncological notes, targeting six primary clinical questions using various prompt strategies.

In addition, a couple of efforts have been made in the domain of radiology report extraction and structuring. Among them, (Steinkamp et al., 2019) proposed a prototype system using traditional ma-

¹The model is made available through this link: [SRI](#)

chine learning techniques to extract useful information from abdominopelvic radiology reports. More recently, (Di Palma et al., 2025) applied LLMs to extract key radiological features specifically from prostate MRI reports.

Our work however, focuses on extracting and structuring doctor notes based on the MRI reports. We extract all clinically relevant information, including details that may not be explicitly stated, resulting in a comprehensive, structured dataset that goes beyond the scope of previous works in both scale and depth. We utilized LLMs in our work due to their high accuracy compared to traditional rule-based or machine learning models. Additionally, they significantly reduce the time and effort required for manual annotation.

By designing diverse prompt strategies, namely, one-question, multi-question, and all-question prompting, combined with various prompting settings such as zero-shot, one-shot, and few-shot, we guided the LLMs to simulate structured data from unstructured notes on MRI reports provided by doctors. This approach enabled the creation of a high-quality synthetic dataset that closely mirrors real clinical annotations.

We began our experiments using a basic instruction prompt with the Llama 3.1 8B Instruct model (Grattafiori et al., 2024), which achieved a composite score of **61.16%**. Building upon this baseline, we enhanced our prompt design and expanded the evaluation to include 6 other LLMs. Among them, Qwen3 4B Instruct-2507 (Team, 2025) achieved the highest composite score at **64.49%**, followed closely by the Qwen2.5 7B Instruct at **63.84%**. While DeepSeek R1 Distill Llama 8B (DeepSeek-AI, 2025) and DeepSeek R1 Distill Qwen 7B (DeepSeek-AI, 2025) attained accuracies of **63.01%** and **53.94%**, respectively, Llama 3.2 3B Instruct (Grattafiori et al., 2024) and MMed Llama 3 8B (Qiu et al., 2024) recorded the lowest performance, with accuracies of **35.56%** and **34.86%**.

In our intermediate set of experiments, we enhanced the prompt and applied different prompt strategies as well as the settings on all models. Subsequently, in the final phase, we fine-tuned all models, with Llama 3.1 8B Instruct, achieving the highest composite score of **87.36%**

Thus, through this multi-level experimental setup, we created a structured dataset that can be effectively utilized to train and evaluate information extraction models, particularly for converting unstructured medical text into structured formats.

2. Dataset Creation

A major challenge in building an unstructured-to-structured data extraction system is the lack of publicly available labeled datasets containing doctor

notes on MRI reports and their corresponding structured outputs. We used a total of 630 unstructured doctor notes on MRI reports collected from a large academic health center. The Institutional Review Board (IRB) of the institution granted a waiver of informed consent for the use of de-identified radiology reports.²

To convert these notes into a structured format, we extracted 55 variables from each report using the Llama 3.1 8B Instruct model, guided by a basic instruction prompt developed by the specific requirements provided by the medical team. The resulting structured data was labeled as V1 (Version 1).

To validate the composite score and clinical relevance of the extracted attributes, doctors manually reviewed and corrected the outputs for the first 60 reports. These validated annotations were recorded as V2 (Version 2). When there was a disagreement between the extracted data and the clinicians, we refined the prompts in consultation with them to align with clinical requirements.

Using the V2 data as a reference, we further experimented with various prompting strategies across multiple models to generate structured outputs for an additional 65 reports. This output was labeled as V3 (Version 3).

Subsequently, doctors reviewed and validated the V3 outputs, resulting in the final refined dataset, referred to as Version 4 (V4). In this dataset, the first 60 reports served as the gold standard test set for evaluating the performance of all prompting strategies and models. The remaining 65 reports were used as example data for constructing the various prompt strategies used in the study.

Additionally, we took another 300 samples (from 630 notes) for training the models. This was referred to as Version 5 (V5) and was also validated by the doctors. The initial round of all validations was performed by us, followed by a second round conducted by the doctors. To facilitate easy validation, we developed a [web application](#) that was used by both the clinicians and us. Below, we describe the various prompt-tuning strategies we employed to optimally extract all variables for the construction of the dataset.

2.1. Prompt-Strategies

We tested three main types of prompt settings: zero-shot, one-shot, and few-shot. Within each of these categories, we further experimented with three prompt strategies: one-question prompt, multi-question prompt, and all-question prompt, to extract the required variables from the reports.

²However, due to privacy concerns, we are not releasing the dataset as a part of this paper.

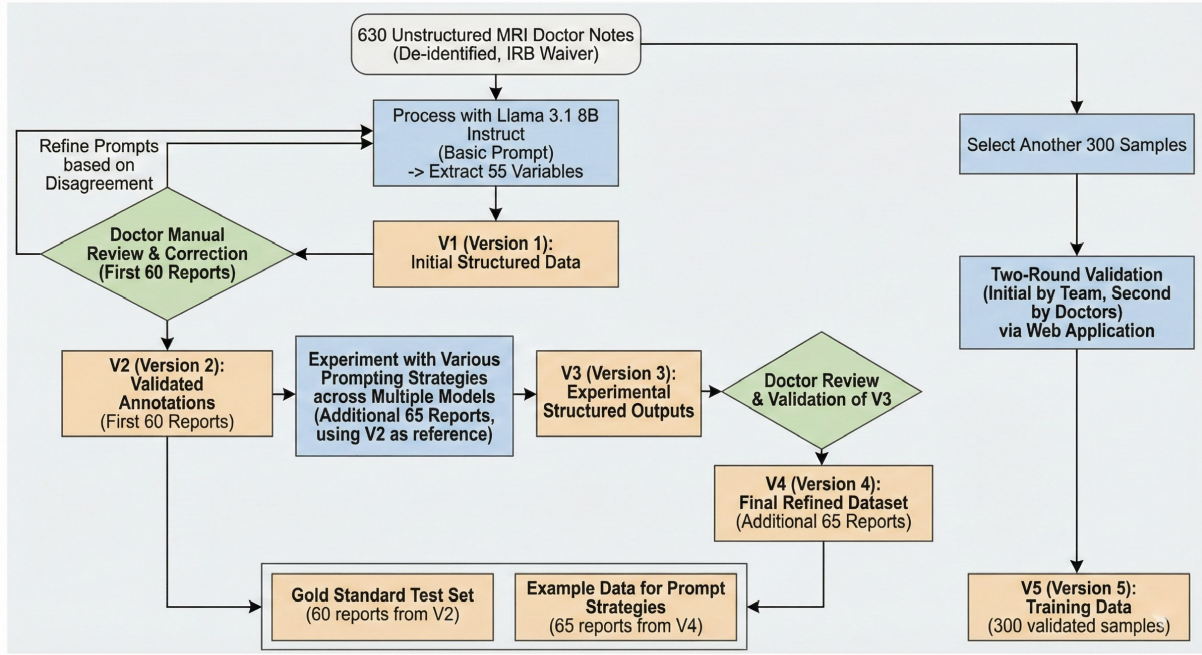


Figure 1: Dataset creation Flow-Chart

Dataset version	Size
Version 1 (V1)	630
Version 2 (V2)	60
Version 3 (V3)	65
Version 4 (V4)	125
Version 5 (V5)	300

Table 1: Summary of dataset versions and corresponding sample sizes

Initially, we planned to use the LLM to generate responses for all 55 variables by asking one question per prompt (shown in [Baseline Prompt \(Appendix A\)](#)). The idea was that generating the answers first using the LLM would make the annotation process easier and more efficient. For this, we used the Llama 3.1 8B Instruct model to generate the initial version of the data, which we referred to as V1. Then, doctors reviewed and refined only 60 reports in the initial data, resulting in a validated version referred to as V2. When we compared V1 with V2, the overall match composite score was **61.16%**. Then we experimented with multiple models using the same baseline prompt.

In this comparison, we observed several issues in the model-generated responses. The model often repeated numeric values unnecessarily, marked certain fields as ‘not mentioned’ even though they were clearly present in the report, and provided inconsistent answers for binary questions such as ‘yes’ or ‘no’. We further extended our experiments to 6 other models (results summarized in [Table 2](#)). Across these models, we continued to observe sim-

ilar issues such as repeated numeric values, incorrect ‘not mentioned’ labels, inconsistent answers to binary questions, and hallucinations.

To address these issues and improve the quality of the generated responses, we refined the prompting approach and adopted three different prompt strategies. They are discussed below:

1. **One-question prompt:** refers to a strategy where each of the 55 variables is derived from a separate, individual prompt. In this approach, we created a distinct question for each variable, resulting in one variable per prompt. Each prompt was designed to include all relevant information required to generate an accurate response, including contextual dependencies and supporting details related to that specific variable. This ensured that the model had sufficient context to produce a meaningful and precise answer for each prompt individually. The key distinction between the baseline prompt and the enhanced one-question prompt lies in the level of specificity provided for variable extraction. In the enhanced version, we explicitly defined how each variable should be identified and extracted, including cases where the information was implicitly stated rather than directly mentioned. Example: [Enhanced One Question Prompt \(Appendix A\)](#)
2. **Multi-question prompt:** refers to a strategy where related variables were grouped based on patterns and dependencies observed in the report responses. Instead of querying each variable individually, we combined several logi-

cally connected questions into a single prompt. These groupings were based on the nature of the answers, such as numeric values or binary (yes/no) responses, and their contextual interdependence. For instance, variables that are often mentioned together in the reports or rely on shared information were grouped to ensure the model had a broader context while still keeping the prompt concise. Example: [Multi-Question Prompt \(Appendix A\)](#)

3. **All-question prompt:** involves presenting all 55 questions within a single, comprehensive prompt. Each question in the prompt is designed to extract one specific variable, covering the entire set of 55 variables in one go. In this approach, we provided all the relevant information and instructions in a single prompt to ensure the model had complete context and access to all interdependent data points.

So far, all the prompt strategies were only applied in the zero-shot prompt setting, where we provided only variable extraction-specific instructions without including any examples for inputs or outputs. Therefore, we wanted to experiment with one-shot and few-shot prompt settings to assess the models' ability to generate accurate and relevant responses under varying levels of context and examples.

1. **One-shot setting:** For one-shot setting,
 - in the **one-question prompt** strategy, we provided one example for each prompt where the target variable was explicitly mentioned in the report. This helped the model understand how to extract that specific variable based on a concrete example.
 - in the **multi-question prompt** strategy, each prompt was accompanied by one example in which most of the grouped variables were clearly present in the report. This allowed the model to understand how to handle interdependent variables within a shared context.
 - in the **all-question prompt** strategy, we included a single comprehensive example containing a report where the majority of the 55 variables were explicitly mentioned. This example served as a reference for the model to understand how to extract all variables from a complete and information-rich report.
2. **Few-shot setting:** For the few-shot setting, we provided three examples for each prompt to help the model learn from a variety of scenarios and improve its generalization capability across different levels of report completeness

- In the **one-question prompt** strategy, each prompt included:
 - (a) An example where the target variable was explicitly mentioned in the report.
 - (b) An example where the target variable was not mentioned in the report.
 - (c) An example of the complete radiology report.
- In the **multi-question prompt** strategy, each prompt included:
 - (a) An example in which most of the grouped variables were explicitly mentioned in the report.
 - (b) An example where some of the grouped variables were either not mentioned or marked as not applicable.
 - (c) An example of the complete radiology report.
- In the **all-question prompt** strategy, each prompt included:
 - (a) An example radiology report where most of the 55 variables were mentioned.
 - (b) An example where a subset of variables was either missing or not applicable.
 - (c) An example of the complete radiology report.

The intermediate results are summarized in [Table 3](#).

3. Finetuning

Despite experimenting with different prompt strategies across various prompt settings, the best-performing model, Llama 3.1 8B Instruct, achieved a maximum composite score of only **77.69%**. Therefore, to further improve performance, we proceeded with fine-tuning the models to determine whether additional gains in composite score could be achieved.

The V4 and V5 versions of the dataset, consisting of 125 + 300 notes, were utilized in the fine-tuning of the models. From these combined 425 notes, 345 were allocated for training, 20 for validation, and 60 for testing.

Fine-tuning was performed using each model's LoRA configuration and employed the one-question prompt strategy.

Overall, fine-tuning led to a clear improvement in performance, demonstrating measurable gains in composite score compared to the intermediate experiment. The detailed results are presented in [Table 4](#).

4. Evaluation

To evaluate model performance, we applied 2 different metrics based on the type of the variable. For variables with numerical outputs, we computed an exact match score. The model's prediction for each variable was compared against the corresponding value in the gold-standard data, assigning a score of 1 for an exact match and 0 otherwise. The report-level score was then obtained by summing these scores across all numerical variables and dividing them by the total number of variables. For variables involving multi-label classifications, we calculated the F1 score using the same ground-truth annotations as the reference. We then calculate the average of the exact match and the F1 scores and report it as the **composite score**. Thus, this composite score forms the final result on the basis of which models are compared, and the best-performing model is noted.

$$CS = \frac{1}{2} \left(\frac{1}{N_{\text{num}}} \sum_{i=1}^{N_{\text{num}}} \mathbf{1}(\hat{y}_i = y_i) + \frac{2PR}{P + R} \right)$$

5. Results

We assessed the performance of the LLMs based on the type of prompt provided at each stage. For the baseline experiment, we evaluated the models using a simple instruction prompt (one-question prompt). The results of the baseline evaluation are summarized in Table 2.

In our intermediate experiment, we systematically evaluated the performance of the models across all three prompt strategies: one-question, multi-question, and all-question under zero-shot, one-shot, and few-shot settings. The detailed results of this evaluation are presented in Table 3.

Our intermediate experiment results show that:

- Under zero-shot setting, across all three prompting strategies, the Llama 3.1 8B Instruct model performed better than the other models. Using the one-question, multi-question, and all-question strategies, it achieved 73.19%, 72.96%, and 74.68%, respectively.
- Under one-shot setting, across all three prompting strategies, the Llama 3.1 8B Instruct model performed better than the other models. Using the one-question, multi-question, and all-question strategies, it achieved 73.98%, 71.94%, and 75.56%, respectively. However, Qwen2.5 7B Instruct performed comparably to Llama 3.1 8B Instruct on the one-question prompt under the one-shot setting.
- Under few-shot setting, the Llama 3.1 8B Instruct model outperformed the other models

on the one-question and all-question strategies, achieving 74.91% and 77.69%, respectively. However, the MMed-Llama 3 8B model achieved 72.22%, and Llama 3.1 8B Instruct performed nearly the same as MMed-Llama 3 8B on the multi-question prompt under the few-shot setting.

LlaMA 3.1 8B Instruct achieved the highest composite score in both the baseline and intermediate experiments.

After fine-tuning, Llama 3.1 8B Instruct again outperformed all other models. Using the one-question, multi-question, and all-question strategies, it achieved 85.09%, 86.71%, and 87.36%, respectively. The overall report-level composite score reached **87.36%**, representing an improvement of nearly 18.42% over the baseline results.

Further, we also observed that the hallucination rate after finetuning was 0.

6. Discussion

We examine several errors observed in the models' performance across three different stages: baseline, intermediate, and fine-tuned. In the baseline setting, the models are provided with a simple instruction prompt and asked to extract both explicit and implicit variables. In the intermediate stage, the prompt is enhanced with more detailed guidelines along with examples to support variable extraction. And in the fine-tuned stage, the enhanced prompt is retained but the examples are removed, and the models are asked to extract the variables using only the refined instructions.

6.1. Baseline Stage

1. We observed that in the single-question prompt, the Llama models frequently returned "not mentioned" for certain variables even when answers were explicitly stated. For example, although the report explicitly mentioned DWI (Diffusion-weighted imaging) as 'dwi – restricted', the model still responded with 'not mentioned'.
2. The models also incorrectly responded "yes" for variables that were not affected by the tumour. For example, the report states: 'PROSTATE AND SEMINAL VESICLES – Normal,' indicating that the prostate and seminal vesicles are not affected by the tumour; however, the model incorrectly returned the response 'yes', implying involvement.
3. For numeric variables such as distance from the anal verge and distance from the ano-rectal junction, the models often repeated the 'tumour length' value in unrelated fields. For

Model	Single	Intermediate	All
Llama 3.2 3B Instruct	35.56%	47.78%	45.23%
Llama 3.1 8B Instruct	61.16%	64.81%	68.94%
MMed Llama 3 8B	34.86%	56.71%	51.25%
Qwen3 4B Instruct-2507	64.49%	65.79%	62.50%
Qwen2.5 7B Instruct	63.84%	67.96%	63.01%
DeepSeek R1 Distill Llama 8B	63.01%	61.02%	67.04%
DeepSeek R1 Distill Qwen 7B	53.94%	55.14%	43.47%

Table 2: Baseline Result: Comparison of Model composite score on baseline prompts

Prompt strategy	Composite Score		
	zero-shot	one-shot	few-shot
Llama 3.2 3B Instruct			
One-question	45.97%	44.31%	42.18%
multi-questions	34.68%	31.94%	35.00%
All-Questions	39.12%	33.38%	31.62%
Llama 3.1 8B Instruct			
One-question	73.19%	73.98%	74.91%
multi-questions	72.96%	71.94%	72.18%
All-Questions	74.68%	75.56%	77.69%
MMed-Llama 3 8B			
One-question	54.81%	51.94%	49.31%
multi-questions	71.02%	70.09%	72.22%
All-Questions	70.00%	71.44%	72.36%
Qwen 4B			
One-question	59.44%	58.94%	54.77%
multi-questions	68.15%	69.44%	71.16%
All-Questions	65.69%	63.10%	65.69%
Qwen 7B			
One-question	56.39%	58.89%	54.17%
multi-questions	68.80%	71.02%	71.81%
All-Questions	64.81%	64.63%	62.08%
DeepSeek Llama 8B			
One-question	69.54%	67.87%	70.46%
multi-questions	71.57%	67.73%	70.19%
All-Questions	70.23%	71.25%	69.17%
DeepSeek Qwen 7B			
One-question	53.94%	54.86%	53.61%
multi-questions	61.02%	63.01%	62.78%
All-Questions	41.94%	39.86%	44.49%

Table 3: Intermediate Result: Comparison of Model composite score on one-question, multi-question, and all-question under zero-shot, one-shot, and few-shot settings.

- example, although the report explicitly stated 'tumour length: 12 mm', the model incorrectly returned 'distance from the anal verge: 12 mm' and 'distance from the ano-rectal junction: 12 mm'.
- In the multi-question prompt, the models returned "not mentioned" for some numeric variables while the correct answer should have been "nil significant."
- Similar to the single-question prompting, the models also frequently returned "not mentioned" for most variables under the all-question prompting approach.
- In both the single-question and multi-question prompts, the Qwen models frequently returned "not mentioned" for gender-related variables and stage-related variables. However, they produced appropriate outputs for numeric variables, correctly reporting the values along with

Model	Single	Multi	All
Llama 3.2 3B Instruct	45.42%	40.37%	47.59%
Llama 3.1 8B Instruct	85.09%	86.71%	87.36%
MMed-Llama 3 8B	79.26%	81.30%	80.79%
Qwen 3 4B Instruct	56.71%	65.79%	64.17%
Qwen 2.5 7B Instruct	62.31%	68.84%	70.32%
DeepSeek R1 Distill Llama 8B	73.89%	76.20%	75.93%
DeepSeek R1 Distill Qwen 7B	71.62%	74.58%	73.38%

Table 4: Finetuned Result: Comparison of Model composite score after fine-tuning.

their respective metrics.

- On the other hand, DeepSeek models, particularly the DeepSeek version of the Llama model performed nearly as well as the Llama 3.1 8B Instruct model. However, both models occasionally produced Chinese characters in the output. Additionally, they often converted numbers and their corresponding units, especially for tumour length; for example, converting 1.2 cm to 11 mm or 125 mm to 12.5 cm. It was also observed that the DeepSeek models exhibited ambiguity when distinguishing among the following variables: tumour length, distance from the anal verge, and distance from the ano-rectal junction. They often returned the tumour length when asked for the distance from the anal verge or the distance from the ano-rectal junction, even when these values were explicitly stated. The models also consistently confused report types, including the first MRI report, restaging MRI, and others. For example, when the report was a first MRI, the other report types should have been marked as "no," but the models returned "yes" or "not mentioned." Similarly, when the report was a post-treatment MRI, they sometimes incorrectly marked it as a "first MRI report".

These were some major reasons for a lower score in the baseline stage.

6.2. Intermediate

- Although the prompts were enhanced, the Llama 3.1 3B Instruct model did not demonstrate any significant improvement across the three prompting settings (zero-shot, one-shot, and few-shot). In contrast, the MMed-LLaMA 8B model performed comparably to the Llama 3.1 8B Instruct model (which performed the best). The primary issue observed was that the models consistently returned "not mentioned" for MRI report stages. A similar pattern was identified for Tumour Regression Grade

and treatment response, where the models frequently defaulted to "not mentioned."

- Qwen performed better than the other models including Llama 3.1 8B Instruct model when using the multi-question prompt. It produced more accurate outputs, particularly for tumour staging variables such as T-stage, N-stage, and M-stage. Additionally, it correctly identified CRM involvement, EMVI, tumour deposit (TD), and T4a status.
- DeepSeek Llama performed well overall. However, it frequently generated unnecessary additional text, including explanations and repetitive content. Further, DeepSeek Qwen often reproduced the example outputs provided in the one-shot and few-shot prompts, rather than generating responses based on the actual input. Despite major improvements, the models still demonstrated confusion among the following variables: tumour length, distance from the anal verge, and distance from the ano-rectal junction. These values were interchanged, although they were explicitly stated in the report. Furthermore, when lymphadenopathy was clearly reported as "nil significant," the models should have returned "nil significant" for internal iliac, obturator, inguinal, external iliac, common iliac, and para-aortic nodes. Instead, they often responded with "not mentioned" or, at times, generated incorrect random values from the examples.

6.3. Fine-tuned

- After fine-tuning, not all the models showed substantial improvement. Although MMed-LLaMA 8B improved to some extent, its performance was still inferior to Llama 3.1 8B Instruct. Infact, Qwen showed a decline in output quality after fine-tuning, frequently returning "not mentioned" for most variables. Although the multi-question prompt continued to produce relatively good outputs, it did not demonstrate

significant improvement compared to the intermediate stage. In contrast, the all-question prompt showed slight improvement after fine-tuning, particularly in generating accurate numeric values.

2. Llama 3.1 8B Instruct produced the most accurate outputs across single-question, multi-question, and all-question prompting strategies. However, all the versions of Llama models still returned “not mentioned” for certain variables that were explicitly stated in the text. We infer that this issue may be related to training data imbalance. In our dataset, some variables such as levator ani, piriformis, and obturator (internus and externus) were predominantly labeled as “not mentioned” and occurred only rarely. Nevertheless, the models appeared to default to “not mentioned” even when these anatomical structures were clearly described in the report.
3. It was also observed that the DeepSeek models produced more accurate outputs for anal sphincter complex variables when compared to the intermediate stage. Moreover, the generation of unnecessary extra content was reduced drastically, and the responses became more concise, containing only minimal additional text. Metric and unit conversion errors were corrected, and hallucinations were largely eliminated. They also performed poorly in identifying post-treatment changes, such as hypointense changes, residual tumour with the same signal as the first MRI, and mucin reaction with a T2 hyperintense focus within a hypointense post-treatment change.

7. Conclusion

This study demonstrates the effectiveness of LLMs, particularly LLaMA 3.1 8B Instruct, in extracting structured information from unstructured MRI reports. A key contribution of this work is the creation of a novel dataset comprising of 55 clinically relevant variables, specifically tailored for doctor notes of MRI reports focused on tumors. Through the evaluation of prompt-tuning strategies across three experimental stages, we showcase the potential of LLMs to automate clinical data extraction with a high composite score. This dataset, along with our experimental findings, provides a valuable resource and foundation for future research in medical information extraction and supports the development of more reliable, scalable, and efficient NLP systems in healthcare.

8. Future Work

As a part of future work, we plan to extend the analysis to additional report types, including CT, biopsy, CEA, and surgical histopathology reports. Furthermore, we intend to incorporate a larger number of reports and curate additional cases in which most of the relevant variables are explicitly documented, to improve data balance and overall model performance.

9. Limitations

A significant proportion of the dataset contained variables labeled as “not mentioned,” which introduced class imbalance across the variables in the dataset. Additionally, certain variable outputs were unevenly distributed. For example, within the T-Stage variable, the categories T3B and T4B appeared far more frequently than T0, T1, and T2, further contributing to class imbalance within the dataset. These were the only notable limitations identified.

10. Acknowledgements

We would like to express our sincere gratitude to Aisha Lakhani, Anju John, Joann Martha Mathew, Krishna Bharati, Lakshmi Meenakshi, Mac Win S, Mohanapriya A, Niveda Bharrath, Saloni Yadav, Shobiga Natarajan, and Sneha H, from CMC Vellore, who contributed their time and expertise to the validate the annotated data.

11. Bibliographical References

- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#).
- Luca Di Palma, Fatemeh Darvizeh, Marco Ali, and Deborah Fazzini. 2025. Structured transformation of unstructured prostate mri reports using large language models. *Tomography*, 11(6):69.
- 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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret,

Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj

Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh

- Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madihan Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojuan Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. [The llama 3 herd of models](#).
- Ray Hashman Hashemi, William G Bradley, and Christopher J Lisanti. 2012. *MRI: the basics: The Basics*. Lippincott Williams & Wilkins.
- Ajith John and Rohin Mittal. 2025. [Role of the multidisciplinary team \(mdt\) meetings in rectal cancer management](#). *Journal of Gastrointestinal and Abdominal Radiology*, 08.
- Vasileios Ntinopoulos, Hector Rodriguez Cetina Biefer, Igor Tudorache, Nestoras Papadopoulos, Dragan Odavic, Petar Risteski, Achim Haeussler, and Omer Dzemali. 2025. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. *BMJ Health Care Inform*, 32(1).
- 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](#).
- Ty J Skyles, Isaac J Freeman, Georgewilliam Kalibbala, David Davila-Garcia, Kendall Kiser, Silpa Raju, and Adam Wilcox. 2025. Exploring chatgpt 3.5 for structured data extraction from oncological notes. *AMIA Summits on Translational Science Proceedings*, 2025:518.
- Jackson M Steinkamp, Charles Chambers, Darco Lalevic, Hanna M Zafar, and Tessa S Cook. 2019. Toward complete structured information extraction from radiology reports using machine learning. *Journal of digital imaging*, 32(4):554–564.
- Qwen Team. 2025. [Qwen3 technical report](#).
- Huaiyuan Ying, Hongyi Yuan, Jinsen Lu, Zitian Qu, Yang Zhao, Zhengyun Zhao, Isaac Kohane, Tianxi Cai, and Sheng Yu. 2025. [Genie: Generative note information extraction model for structuring ehr data](#).

A. Prompt Appendix

Baseline Prompt

You are an expert radiology information extraction model. Respond only in JSON.
Extract the n-stage from the radiology report.

OUTPUT FORMAT:

{'value': 'N0, N1, N1c, N2, or not mentioned'}

REPORT: {REPORT}

Enhanced One Question Prompt

You are an expert radiology information extraction model. Your task is to extract the current length of the tumour from the given radiology report text based on the detailed instructions below and only respond in JSON format.

Instructions:

You must follow these exact guidelines to generate answer:

1. If the radiology report clearly states phrases such as:
 - "No foci of T2 intermediate signal intensity or restricted diffusion in this segment"
 - "Likely post-RT changes"
 - "Post-treatment changes in the low rectum with no evidence of residual/recurrent tumor"→ This indicates a complete radiologic response. In this case, return: "Not Relevant"
2. If tumour length is explicitly mentioned using terms such as:
 - "length"
 - "tumour size"
 - "tumour length"→ Extract the numeric value of length of the tumour and its associated unit, ensuring there is a space between the number and the unit (e.g., "46 mm", "3.2 cm").
→ Always use numerals instead of words (e.g., "4 mm", not "four mm").
3. If the report is a post-treatment complete radiology report, and it only mentions scar size but not tumour size, extract the scar size as the current tumour length.
4. If neither tumour size nor scar size is mentioned, but the report qualifies as complete (based on the text), return: "Not Relevant"
5. If the tumour length is not mentioned, and the report is not complete, return: "Not Mentioned"

Multi Question Prompt

You are an information extraction system that can extract information from a given text and only respond in JSON format.

for 'crm_or_mrf_involvement':

1. If crm or mrf involvement is explicitly mentioned using terms such as:
 - "CRM - involved"
 - "CRM - 0 mm"
 - "mesorectal fascia - 9 mm"
2. return "Yes" if the distance to the mesorectal fascia is explicitly reported as less than 1 mm.
3. return "No" if the distance is explicitly reported as greater than or equal to 1 mm.
4. Return "Not Mentioned" if the distance to the mesorectal fascia is not mentioned in the text.

for 'emvi':

1. If EMVI is explicitly mentioned using terms such as:
 - "EMVI - present"
 - "EMVI - nil"
 - "EMVI - absent"
2. return "Yes" if EMVI is present and not mentioned as fibrosed, healed.

3. return "No" if EMVI is mentioned as fibrosed, healed, resolved, absent, or nil.
 4. return "Not Mentioned" if EMVI is not explicitly mentioned.
- for 'tumour_deposit':
1. return "Yes", if the report explicitly mentions the presence of tumor deposits, or uses terms like:
 - "tumor deposit (TD) - present"
 - "ENTD (extranodal tumor deposit) - present"
 2. return "No" if the report explicitly mentions the absence of tumor deposits, or uses terms like:
 - "tumor deposit (TD) - absent"
 - "ENTD (extranodal tumor deposit) - absent"
 3. return 'Not Relevant' if the input text is a complete radiology report.
 4. return "Not Mentioned" if tumor deposit (TD) or ENTD (extranodal tumor deposit) are not explicitly mentioned.
- for 'is_t4a':
1. return "Yes", if tumor invasion of the pelvic peritoneal reflection is explicitly mentioned.
 2. return "No", if the pelvic peritoneal reflection is not involved, or confirms that the tumor is below the peritoneal reflection.
 3. return "Not Mentioned", if the pelvic peritoneal reflection is not mentioned.

B. Example Appendix

Semi-Structured or Unstructured Report Example

MRI ABDOMEN AND PELVIS

FINDINGS:

RECTUM: Residual circumferential thickening of the low rectum with peri-rectal fat stranding , inferiorly extending to anorectal junction. Shows focus of T2 intermediate signal intensity with small focus of restricted diffusion in left lateral wall

- length - 15mm vs 20 mm previously
- extramural spread - 2.4 mm
- CRM - involved at 12 Oclock position by tumor
- CRM - 0 mm - EMVI - nil
- highest margin of the tumor is below the peritoneal reflection
- no infiltration of puborectalis or anal sphincter complex
- no infiltration of vagina, uterus, cervix or bladder
- mesorectal nodes: nil significant

LYMPHADENOPATHY: para-aortic - Nil significant internal iliac - Nil significant external iliac - Nil significant inguinal - Nil significant

LIVER - No focal lesions in the liver, no IHBRD. SPLEEN - Normal GB - Distended, no wall thickening or calculi. CBD is of normal calibre. PANCREAS - Normal ADRENALS - Normal KIDNEYS - Normal BOWEL, MESENTERY, OMENTUM - Rectum as described.

Rest nil significant FLUID - nil

BLADDER - Minimally distended, normal

UTERUS - multiple fibroids, largest 15x13mm in the anterior wall

OVARIES - Nil significant

INGUINAL ORIFICES - Nil significant

ABDOMINAL WALL - Divarication of rectus noted.

BLOOD VESSELS -Nil significant

VISUALISED LUNG BASES -Nil significant

VISUALISED BONES -Nil significant

IMPRESSION: 70 year old lady with carcinoma rectum, post treatment on follow up, MRI shows

1. Residual circumferential thickening of the low rectum, extending to anorectal junction with T2 intermediate signal intensity and small focus of restricted diffusion CRM =0 EMVI - nil
- 2.No significant lymphadenopathy.

No significant change from previous MRI dated Feb 2023.

Suggest clinical evaluation and close follow up.

Structured Report Example

```
{
  "report_type": "MRI",
  "bio": {
    "age": "70",
    "gender": "Female"
  },
  "mri_numeric": {
    "current_length_of_tumour": "15 mm",
    "distance_from_anal_verge": "Not Mentioned",
    "distance_from_anorectal_junction": "0 mm",
    "extramural_spread_size": "2.4 mm",
    "mr_crm_distance": "0 mm",
    "number_of_mesorectal_nodes": "Nil Significant",
    "internal_iliac_nodes": "Nil Significant",
    "size_of_obturator_nodes": "Not Mentioned",
    "size_of_inguinal": "Nil Significant",
    "size_of_external_iliac": "Nil Significant",
    "size_of_common_iliac": "Not Mentioned",
    "size_of_para_aortic": "Nil Significant"
  },
  "mri_stage": {
    "is_first_mri_report": "No",
    "is_restaging_mri_report": "Yes",
    "is_mri_report_after_neoadjuvant": "Not Mentioned",
    "is_post_treatment_mri_report": "Yes",
    "is_post_operative_mri": "Not Mentioned"
  },
  "location_of_the_tumour": "low",
  "t2_signal_intensity": "Intermediate",
  "dwi": "Restricted diffusion",
  "morphology": "Not Mentioned",
  "post_treatment_change": {
    "is_thick_t2_hypointense_band": "Not Mentioned",
    "is_thin_t2_hypointense_band": "Not Mentioned",
    "is_residual_tumor_as_first_mri": "Not Mentioned",
    "is_mucin_reaction_t2_hyper_hypo_post_treatment": "Not Mentioned"
  },
  "rectal_perforation": {
    "obstruction": "Not Mentioned",
    "perforation": "Not Mentioned"
  },
  "radial_extent": "Annular or Circumferential",
  "crm_or_mrf_involvement": "Yes",
  "emvi": "No",
  "tumour_deposit": "Not Mentioned",
  "is_t4a": "No",
  "adjacent_structures_t4b": {
    "female": {
      "uterus": "No",
      "vagina": "No",
      "ovaries": "No"
    },
    "male": {
      "prostate": "Not Applicable",
      "seminal_vesicles": "Not Applicable"
    },
    "male_and_female": {
      "levator_ani": "Not Mentioned",
      "puborectalis": "No",
      "piriformis": "Not Mentioned",
      "obturator_internus": "Not Mentioned",
      "obturator_externus": "Not Mentioned"
    }
  },
  "anal_sphincter_complex": {
    "internal_sphincter": "No",
    "t4b": {
      "external_sphincter": "No",
      "inter_sphincteric_plane": "No",
      "ischiorectal_foss": "No",
      "fistula_in_ano": "No"
    }
  },
  "t_stage": "T3b",
  "n_stage": "N0",
  "m_stage": "M0",
  "mr_trg": "MR TRG 2",
  "response": "Near Complete"
}
```

Useful to Whom? A Persona-Driven Evaluation of Knowledge-Adapted Health Question Reformulation via LLM Simulation

Jooyeon Lee, Luan Huy Pham, Özlem Uzuner

George Mason University
4400 University Dr, Fairfax, Virginia, United States of America
{jlee252, lpham6, ouzuner}@gmu.edu

Abstract

Automatic metrics such as F1 and BERTScore are often insufficient for evaluating user-centric generative tasks like Consumer Health Question (CHQ) reformulation. A high F1-score may not correlate with user satisfaction, especially when the user's knowledge level (UKL) dictates their needs. We propose a robust, Persona-Driven Evaluation Framework (PDEF), grounded in cognitive science and health literacy literature, to measure persona-specific UCQ. This framework assesses reformulations from the perspectives of a 'Layperson' (requiring foundational context) and an 'Expert' (requiring efficient, precise answers). We apply this framework to a set of reformulated questions generated by LLMs, and test the robustness of our evaluation by using three state-of-the-art LLMs (GPT-4o, Llama 3.3, and Mistral Large) as the evaluators. Our results reveal a significant disconnect between automatic metrics and user-perceived quality: the model with the highest F1-score (0.6134) was *consistently outperformed* in user preference by a Pipelined model, with experts preferring the latter by a statistically significant margin ($p < 0.001$). Furthermore, our persona-driven ablation analysis provides robust evidence that specific architectural components, specifically UKL inference and Entailment logic, are linked to significant gains in persona-driven quality for Layperson cohorts. This work demonstrates the critical need for user-centric evaluation and shows that its findings are generalizable across different LLM architectures.

Keywords: Consumer Health, Question Understanding, LLM-as-a-Judge, Persona-Driven Evaluation

1. Introduction

Consumers increasingly turn to the internet for health information, but are often hindered by a "vocabulary gap" (Zeng et al., 2005; Zeng and Tse, 2006) between their layperson queries and the medical jargon of online resources. Consumer Health Question (CHQ) reformulation aims to bridge this gap by converting ambiguous queries into a set of canonical, answerable questions.

However, evaluating these generative systems, specifically Large Language Model (LLM) architectures designed for query reformulation, is notoriously difficult. While these systems excel at producing fluent text, traditional automatic metrics based on semantic similarity, such as F1-scores, often fail to measure the actual utility of the output for specific user cohorts (Zhang et al., 2020). For instance, a reformulation that is verbose and repetitive might achieve a high F1-score via keyword matching, yet frustrate an expert user seeking a concise, evidence-based data point.

We argue that a "one-size-fits-all" reformulation is suboptimal because it ignores the divergent cognitive needs of users. To address this, we formalize user requirements into two distinct profiles based on domain familiarity and information-seeking goals (Brusilovsky, 2001; Hölscher and Strube, 2000):

Novice (Low UKL): A user who lacks a foundational understanding of the specific medical domain, often characterized by "Low Health Literacy" or limited exposure to technical medical silos (White et al., 2008). This typically includes newly diagnosed patients or concerned caregivers.

- **Information-Seeking Goal:** Orientation and Scaffolding. As established by Chiu et al. (Chiu et al., 2022), laypersons with fewer clinical consultations often struggle with match rates between their queries and professional health information. They require foundational context to bridge the "vocabulary gap" and effectively interpret technical answers (Kuhlthau, 1991).
- **Behavioral Indicator:** Novices derive a Shared Benefit from query normalization but specifically require Augmentation. Unlike experts who prioritize simple summarization, novices prefer reformulations that move beyond paraphrasing toward Entailment, the addition of broader, prerequisite questions that provide the necessary background to understand the primary query (Chen et al., 2021; Vakkari, 1999).

Expert (High UKL): a user with high domain-specific knowledge, such as a clinician, medical researcher, or a "pro-patient" with extensive lived

experience, who has developed high search proficiency through frequent medical consultations (Chiu et al., 2022). These users are characterized by their ability to navigate specialized technical resources like PubMed (National Center for Biotechnology Information (NCBI), 2026) to extract evidence-based data (White et al., 2008; Research, 2025).

- **Information-Seeking Goal: Precision and Extraction.** Experts seek specific, clinically relevant data points to fill narrow information gaps within an existing, robust mental model (Chen et al., 2021).
- **Behavioral Indicator: They prioritize Normalization (de-noising complex narrative queries) but actively reject Augmentation.** Due to the Expertise Reversal Effect (Sweller. et al., 2011; Kalyuga et al., 2003), they perceive foundational "scaffolding" (e.g., definitions) as extraneous cognitive load or "noise" that hinders their retrieval efficiency.

To evaluate these divergent needs, we define User-Centric Quality (UCQ) as the degree to which a reformulated query aligns with the specific cognitive constraints and information-seeking goals of a target persona. Grounded in Saracevic's theory of situational relevance (Saracevic, 1975) and Sweller's Expertise Reversal Effect (Sweller. et al., 2011), this metric prioritizes the functional effectiveness of the output over simple string matching. High UCQ ensures that a novice receives necessary scaffolding while an expert is provided with a concise, high-efficiency path to information.

This theoretical conflict, where the same information serves as a bridge for one user and a barrier for another, necessitates a Persona-Driven Evaluation Framework (PDEF). To quantify the alignment between reformulation strategies and these opposing needs, we established four persona-driven evaluation criteria:

1. **Goal Achievement:** How well the output helps the persona achieve their primary goal (orientation vs. efficiency) (Kuhlthau, 1991; Hölscher and Strube, 2000).
2. **Cognitive Appropriateness:** How well the level of context (added or removed) matches the persona's needs (Sweller. et al., 2011; Norman, 1987).
3. **Clarity & Usability:** How easy the output is to understand and use (iso, 2018; Zarcadoolas et al., 2006).
4. **Perceived Relevance & Utility:** How confident the persona is that the output will lead to useful answers for their specific goal (Saracevic, 1975; Vuong et al., 2019).

Model	Description
<i>Baselines</i>	
B1 (Summ.)	Summarizes the query.
B2 (QR Def.)	Given a simple definition of reformulation.
<i>Context-Aware Frameworks</i>	
SPM (Single-prompt)	Infers UKL, Focus and Type and applies entailment logic.
PM (Multi-prompts)	Infers UKL, Focus and Type with separate prompts, then uses those for input for the prompt applying entailment logic.
<i>Ablation Models (SPM)</i>	
SPM (-UKL)	Removes UKL inference and UKL usage from entailment logic.
SPM (-Entail.)	Removes Entailment logic.

Table 1: Reformulation models and baseline strategies sourced directly from (Lee et al., 2026). We utilize these existing model outputs to validate our PDEF against both context-agnostic and adaptive architectures.

By moving beyond traditional semantic overlap (F1-score) to these theory-grounded metrics, we can precisely identify where current LLMs succeed or fail in meeting subjective user requirements.

In this work, we hypothesize that the most effective reformulation is tailored to the UKL. We test this hypothesis by introducing a PDEF, which we apply to a series of generative systems, detailed in Table 1, originally developed in our prior work (Lee et al., 2026).

These systems represent a spectrum of reformulation strategies: Baselines provide context-agnostic outputs such as simple summarization (B1) or a standard definition of the reformulation task (B2); Context-Aware Frameworks with Single-Prompt Model (SPM) and Pipeline-Model (PM) utilize GPT-4 (OpenAI, 2023) to adaptively infer UKL and apply entailment logic, adding foundational context; and Ablation Models allow us to isolate the specific impact of UKL inference and entailment. By evaluating these diverse strategies through our PDEF, we can quantify which model components drive satisfaction for different types of users.

Our contributions are threefold:

1. We introduce a persona-based evaluation framework grounded in health literacy (Kuhlthau, 1991; Zarcadoolas et al., 2006) and cognitive load theory (Sweller. et al., 2011). This framework is designed to quantify the alignment of reformulations with the distinct information-seeking goals and cognitive constraints of user cohorts by implementing user

needs into measurable persona-driven criteria.

2. We validate our prior F1-score findings (Lee et al., 2026) with a user-centric lens, demonstrating that our persona-based evaluation also finds the Context-Aware Frameworks to be dramatically superior to B1, B2, SPM (-UKL) and SPM (-Entail).
3. We validate the robustness of our framework and findings by running our evaluation on three state-of-the-art LLMs (GPT-4o, Llama 3.3, and Mistral Large), showing that our core conclusions are generalizable and not an artifact of a single model.

2. Related Work

Consumer Health Question (CHQ) Simplification Bridging the “vocabulary gap” between laypeople and medical professionals is a foundational challenge in medical informatics (Zeng et al., 2005; Zeng and Tse, 2006). Early approaches relied on dictionary-based substitution and statistical machine translation to simplify jargon. With the advent of neural generation, recent work has shifted toward abstractive reformulation and summarization (Ben Abacha and Demner-Fushman, 2019; Lee et al., 2026). However, the majority of existing research focuses exclusively on *simplification*, making text readable for low-literacy users. Our work diverges by acknowledging that “simplification” is often detrimental to high-knowledge users (experts), necessitating an adaptive approach that can toggle between scaffolding and precision.

User Knowledge Modeling & Cognitive Load

The distinction between novice and expert information needs is a foundational challenge in information retrieval. For nearly two decades, research has shown that domain expertise fundamentally dictates search interaction patterns. This was pioneered by White et al. (White et al., 2008), who utilized specialized resource visitation (e.g., PubMed) to distinguish between medical experts and non-experts (Novices), finding that experts prioritize technical precision over general orientation.

This trajectory continues into the era of generative AI; recent large-scale telemetry studies from Microsoft Research (Research, 2025) confirm that these expertise levels remain the primary drivers of user satisfaction in chat-based interfaces. Specifically, while novices utilize generative tools for information recall and mental model building (Kuhlthau, 1991), experts demand high-complexity, targeted extraction. Our work implements these established behaviors by testing the Expertise Reversal Effect (Sweller et al., 2011; Kalyuga et al., 2003), a

phenomenon where the foundational “scaffolding” essential for a novice’s understanding becomes extraneous cognitive load that hinders an expert’s efficiency.

Evaluation of Generative Reformulation Evaluating generative text remains an open challenge. Traditional n-gram metrics (BLEU, ROUGE) and embedding-based metrics (BERTScore) (Zhang et al., 2020) measure semantic overlap but fail to capture persona-specific requirements or the cognitive appropriateness of the output. Recently, the “LLM-as-a-Judge” paradigm (Zheng et al., 2023) has emerged as a cost-effective alternative to human annotation, correlating well with human preferences on general tasks. However, most LLM-based evaluations use a single, generic evaluator prompt. In this work, we propose a PDEF, using opposing evaluator personas to explicitly measure the trade-off between accessibility (for laypeople) and efficiency (for experts).

3. Methodology

3.1. Experimental Setup

To validate the PDEF, we utilize the Consumer Health Question (CHQ) dataset established in (Lee et al., 2026). This corpus consists of 273 complex, multi-intent queries sourced from the NIH Genetic and Rare Diseases (GARD) Information Center. Each query is paired with expert-curated reformulations that serve as the ground truth for traditional semantic evaluation. Our experiments are conducted on a controlled testbed of 1,638 reformulated outputs (273 queries $\times$ 6 systems). These outputs were generated by the six architectures described in the Introduction (Table 1), ranging from context-agnostic baselines to adaptive, context-aware frameworks. By using this fixed set of outputs, we isolate the PDEF’s ability to measure persona-driven utility across varying levels of architectural complexity.

3.2. Automatic Evaluation Baseline

Following the methodology of our prior work (Lee et al., 2026), we first evaluated all models using a redundancy-aware semantic F1-score (RA-Sem-F1). This metric computes a set-based F1-score by matching generated questions against a gold-standard set using specialized medical embeddings: **BGE-Medical** (Hükmet, 2024) (0.67 similarity threshold) and **BioLORD** (Rémy et al., 2022) (0.90 threshold for deduplication). As shown in Table 2, these semantic metrics establish that the SPM is the top-performing model under traditional evaluation ($F1 = 0.6134$). This serves as the baseline context for our current study, allowing us to

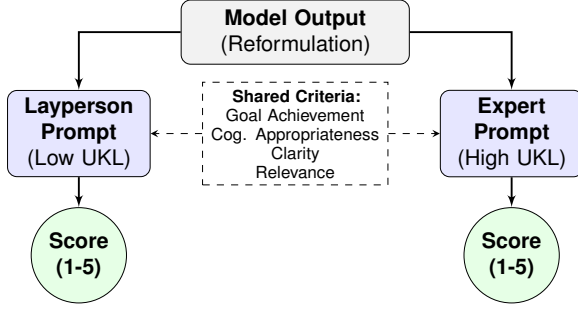


Figure 1: The PDEF. The model output is evaluated by two distinct **Persona Prompts** (Layperson vs. Expert). Both prompts utilize the same **Shared Criteria** to generate a evaluation score.

compare traditional performance against persona-specific utility.

3.3. PDEF Implementation

To strictly measure UCQ, we propose the PDEF, illustrated in Figure 1. We developed a structured evaluation pipeline using LLMs-as-judges (Zheng et al., 2023) to process the 1,638 reformulated outputs through the following procedure:

1. **Prompt Construction:** For every query-output pair, we constructed a specific evaluation prompt containing the original user query, the model’s reformulation, and a detailed Persona Profile (Layperson or Expert).
2. **Scoring Rubric:** Each judge was provided with the four Persona-Driven Criteria established in the Introduction. The judges were instructed to provide a Likert score (1–5) for each criterion, followed by a brief natural language justification to ensure reasoning consistency.
3. **Inference Parameters:** To ensure stable and reproducible scoring, we set the sampling temperature to $\tau = 0.0$ for all three LLM judges (GPT-4o, Llama 3.3, and Mistral Large), minimizing stochastic variance in the evaluation results.
4. **Data Aggregation:** We collected a total of 9,828 individual evaluations (1,638 outputs $\times$ 2 personas $\times$ 3 judges). These scores were then averaged across the four criteria to calculate the Mean Core Criteria score for each model-persona pair.

3.3.1. LLM Judge Selection and Statistical Validation

A key critique of the “LLM-as-a-judge” paradigm is the potential for Self-Preference Bias, where a

Method	Recall	Precision	F1-Score
<i>Baselines</i>			
B1 (Summ.)	0.5353	0.6507	0.5874
B2 (QR Def.)	0.6037	0.5669	0.5847
<i>Adaptive Models</i>			
PM	0.5399	0.6730	0.5992
SPM	0.5544	0.6866	0.6134
<i>Ablations (Single-Prompt)</i>			
SPM (-UKL)	0.4659	0.6710	0.5500
SPM (-Entail.)	0.4703	0.7090	0.5655

Table 2: Performance of reformulation models using traditional semantic metrics (Lee et al., 2026). Key Highlight: The SPM model significantly outperforms context-agnostic baselines in F1-score, establishing it as the benchmark for our subsequent persona-driven analysis.

model preferentially rates outputs generated by itself or similar architectures (Zheng et al., 2023). To mitigate this and ensure our findings are generalizable, we selected three distinct SOTA models to serve as our evaluators, specifically chosen to represent architectural and developer diversity.

We selected **GPT-4o** (OpenAI, 2024) as the industry benchmark. As a closed-source model utilizing a Mixture of Experts (MoE) architecture (Shazeer et al., 2017), GPT-4o represents the current state-of-the-art. Any modern evaluation framework must be validated against this baseline. To ensure architectural diversity, we included **Llama 3.3** (Meta AI, 2024). Unlike MoE models, Llama utilizes a dense architecture where every parameter is activated for every token. This inclusion is critical for validity; by demonstrating agreement between sparse MoE models and dense models, we control for artifacts that might be specific to sparse activation patterns. Finally, we selected **Mistral Large** (AI, 2024) to represent a competitive alternative MoE architecture. Including Mistral allows us to test if our findings hold across different implementations of similar architectural philosophies, specifically comparing GPT-4o against Mistral.

To determine the statistical significance of the performance differences measured by these evaluators, we performed two-tailed paired t-tests ($\alpha = 0.05$) following the methodology described by Montgomery and Runger (Montgomery and Runger, 2003). This statistical validation ensures that the observed utility gains for Laypersons and the consistent penalties recorded in the Expert scenarios are mathematically significant and not due to random variance in the LLM judges’ outputs.

Scenario	Comparison	Mistral	Llama 3.3	GPT-4o
Layperson (UKL LOW)	vs. B1 (Summ.)	+0.122	-0.001	+0.296
	vs. B2 (QR Def.)	-0.082	-0.106	-0.004
	vs. -UKL	+0.641	+0.663	+0.800
	vs. -Entail.	+0.553	+0.532	+0.750
	vs. -F/T	+0.101	+0.060	+0.111
Expert (UKL HIGH)	vs. B1 (Summ.)	+0.033	-0.107	-0.037
	vs. B2 (QR Def.)	+0.846	+1.013	+0.745
	vs. -UKL	-0.284	-0.271	-0.312
	vs. -Entail.	-0.241	-0.281	-0.300
	vs. -F/T	+0.226	+0.120	+0.092

Table 3: Comparative UCQ scores across state-of-the-art LLM Evaluators. Our core findings (in bold) are stable across all three evaluators.

4. Results and Analysis

4.1. Automatic Metric Results

The automatic evaluation results (from our prior work) are shown in Table 2. Based on this evaluation, the **SPM** achieves the highest F1-score (0.6134). The ablation study confirms that removing the **UKL** (0.5500) or **Entailment** (0.5655) components causes a significant drop in F1-score.

4.2. Agreement Between LLM Judges

We define consensus as the directional agreement across three independent LLM judges (GPT-4o, Llama 3.3, and Mistral Large). The values in Figure 2 represent the Mean Core Criteria Delta (Δ), calculated as: $\Delta = \text{Score}_{\text{SPM}} - \text{Score}_{\text{Comparison Model}}$. This metric quantifies the average difference on a 1–5 Likert scale across the four evaluation criteria. A positive (green) delta indicates the SPM provided higher UCQ, while a negative (red) delta indicates the comparison model was preferred by that persona. Utilizing deltas allows us to isolate the specific impact of architectural features like Entailment while controlling for baseline model performance.

A notable trend in this study is the level of agreement across the three LLM judges. As shown in Table 3, GPT-4o, Llama 3.3, and Mistral Large all independently reached similar conclusions: the SPM is preferred for Laypersons, while the "Expert's Dilemma" (the penalization of extra context) is observed across all evaluators. This consistency suggests that the PDEF produces stable results that do not depend on which specific model is performing the scoring.

4.3. PDEF Results

The Persona-Driven Evaluation reveals opposing utility requirements between user cohorts. Figure 2 illustrates the consensus across all three LLM evaluators (GPT-4o, Llama 3.3, and Mistral Large). While the SPM significantly dominates all baselines and ablations for the **Layperson** persona,

consistent green across evaluators, we observe a definitive **Expert's Dilemma**: expert personas consistently prefer the ablated models that lack UKL and Entailment logic as indicated by the red delta blocks.

4.4. Analysis of Key Scenarios

We analyze how the SPM performs against the baselines in the two scenarios it was designed to handle, using the Mistral results.

Layperson (UKL LOW): In this scenario, the SPM, with a mean score of 4.95, **significantly outperforms B1**, which achieved a mean of 4.76 ($\Delta=+0.19$, $p=0.00147$). This confirms our hypothesis: for novices, summarization is the wrong approach, and the model's ability to add foundational context is critical. Against B2, the B2 model is slightly preferred on most metrics ($p=0.105$). However, the SPM achieves a significantly higher Clarity & Usability score ($\Delta=+0.05$, $p=0.0447$). This suggests that while both models provide appropriate context, the SPM's output is clearer.

Expert (UKL HIGH): Here, the SPM mean of 4.37 dramatically outperforms the B2 mean of 3.53 ($\Delta=+0.84$, $p<0.001$). This is our core hypothesis validated. The Expert persona, valuing efficiency, strongly prefers the SPM's concise output. This is most visible in the 'Cognitive Appropriateness' score, which has a massive delta of **+1.06** ($p<0.001$), proving the expert rejects the "extraneous cognitive load" imposed by B2's basic questions.

4.5. PDEF Ablation Analysis

The ablation results, summarized across all evaluators in Figure 2, causally link the model's performance to its specific prompt components.

Layperson (UKL LOW): The full SPM **dominates** both the SPM (-UKL) and SPM (-Entail.) ablations. As shown in the heatmap, all three evaluators record massive positive deltas (e.g., Mistral $\Delta=+0.64$ for -UKL and $+0.55$ for -Entail.). This provides robust evidence that the UKL and Entailment logic are the primary drivers of success for novice users; removing them causes a collapse in user-perceived utility.

In contrast, the impact of removing Focus/Type structuring (-F/T) is significantly smaller ($\Delta=+0.10$). This positions structural deconstruction as a "nice-to-have" for novices, whereas contextual augmentation is "essential."

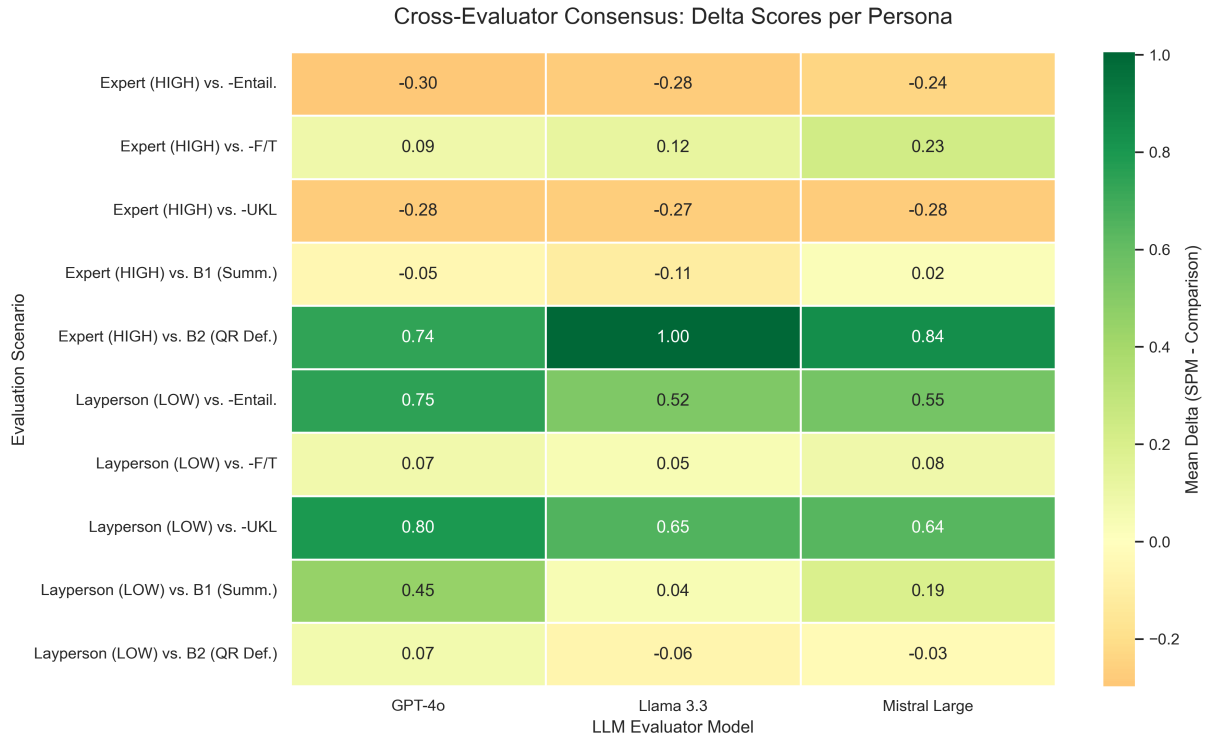


Figure 2: Cross-Evaluator Consensus: Delta scores (SPM - Comparison Model) across three state-of-the-art LLMs. **Green (positive)** indicates the SPM outperformed the comparison; **Red (negative)** indicates the comparison was preferred. Note the consistent **Expert’s Dilemma**: Experts across all evaluators penalize the UKL/Entailment logic that novices require.

Expert (UKL HIGH): We observe a definitive reversal for the expert persona. The **ablated models (-UKL and -Entail.) are significantly preferred** over the full SPM (Mistral $\Delta=-0.28$ and -0.24 , respectively). This is the quantitative signature of the “Expert’s Dilemma”: experts penalize the inclusion of the very components that the laypersons require.

However, the SPM (-F/T) ablation reveals a critical nuance: removing the structural logic **significantly hurts** expert satisfaction ($\Delta = +0.23$, $p < 0.05$). This confirms that experts desire query deconstruction (normalization) but are actively hindered by foundational scaffolding (augmentation).

4.6. SPM vs. PM Analysis

The comparison between architectures reveals a final architectural insight. For the **Layperson**, the PM and SPM remain in a statistical tie ($\Delta=-0.02$, $p > 0.05$). However, for the **Expert**, the **PM consistently and significantly outperforms the SPM** across all evaluators (Mistral $\Delta=-0.35$, $p < 0.001$). This suggests that a decoupled, modular pipeline is more robust at maintaining precision for high-knowledge users than a single-pass inference model.

5. Discussion

Our results highlight five key findings.

5.1. The “Expert’s Dilemma”: A Design Trade-Off

Our analysis reveals a critical system design tension we term the “Expert’s Dilemma.” This refers to the conflict between an expert’s functional need for query deconstruction and their cognitive intolerance for the added context that typically accompanies it.

Quantitatively, our ablation study for Experts showed that removing the contextual logic (-UKL and -Entailment) actually *improved* performance. This, combined with the statistical tie between the SPM and the simple Summarization baseline (B1), suggests that for high-knowledge users, a “less is more” strategy is optimal. However, the -F/T ablation reveals a critical nuance. While Experts penalize the *contextual* logic of UKL, they significantly *benefit* from the *structural* logic of F/T (Focus/Type) deconstruction. This suggests the ideal expert model is not merely a summarizer, but a precise query deconstructor that adds no new foundational context.

To understand the mechanism behind these quantitative results, we qualitatively analyzed the

Input Query	Reformulation (SPM)	Evaluator Reactions (Mistral)
"My son presented with a sudden onset of partial complex seizures... I would like to know if this disease has an affect on a person's appetite or weight?"	1. What is neurocysticercosis? 2. How is neurocysticercosis diagnosed? 3. Does neurocysticercosis impact a person's appetite? 4. Is weight change a potential symptom...?	Layperson (Score: 5/5): "This list helps me understand neurocysticercosis from the basics... making it easy to grasp the topic step-by-step." Expert (Score: 2/5): "Includes too many basic and redundant questions that waste my time, making it inefficient for my goal..."

Table 4: Qualitative comparison showing the "Expert's Dilemma." The Layperson values the scaffolding (Question 1), while the Expert penalizes it as noise.

evaluators' textual justifications. Table 4 presents a representative example (Neurocysticercosis query) that illustrates this divergence. The Layperson persona assigned the reformulated list a perfect score of 5/5, explicitly citing the inclusion of the definition ("What is neurocysticercosis?") as "essential scaffolding." In sharp contrast, the Expert persona penalized the exact same reformulation with a score of 2/5, describing the foundational questions as "redundant" and noting that they "waste my time."

This qualitative evidence confirms our hypothesis: the *UKL* component improves performance for novices by fulfilling their need for orientation, but simultaneously degrades performance for experts by increasing extraneous cognitive load.

5.2. The Tension Between Normalization and Augmentation

Our qualitative analysis indicates that the model performs two distinct functions: *Normalization*, which decomposes CHQs into clean questions, and *Augmentation*, which adds foundational definitions.

Shared Benefit: Normalization. Both personas benefit significantly when the model strips away "narrative noise." For instance, in a query regarding MELAS Syndrome, the user originally provided a dense paragraph detailing their mother's autopsy report and their sister's fears about migraines. The model reduced this entire narrative to four canonical questions (e.g., "*Is MELAS syndrome inherited?*", "*How is it diagnosed?*"). Both the Layperson and Expert evaluators rated this reduction 5/5. The Expert explicitly noted that stripping the "*irrelevant family history*" was a key factor in the high score, confirming that "clinical de-noising" is a universal quality metric that reduces cognitive load for novices while increasing efficiency for experts.

Divergent Preference: Augmentation. The conflict arises solely from *Augmentation*. While the Layperson persona explicitly praises the addition of

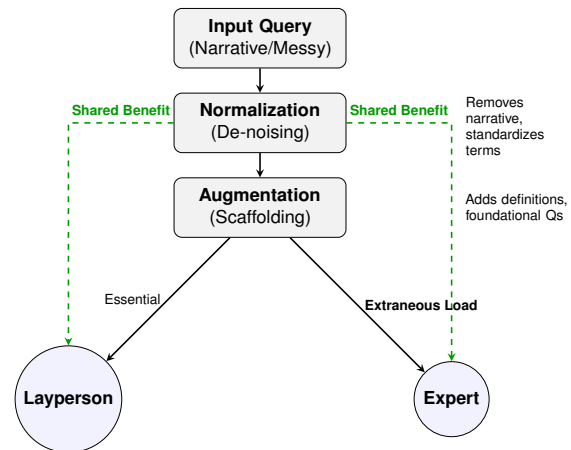


Figure 3: The Mechanism of the Expert's Dilemma. Both personas derive a **Shared Benefit** from Normalization (cleaning the query). The divergence occurs at **Augmentation**: while essential for the Layperson, it introduces extraneous load that the Expert actively penalizes.

definitions (e.g., "What is neurocysticercosis?") as "essential scaffolding" for building a mental model from scratch, the Expert persona penalizes this exact same content as "redundant" and "a waste of time."

This clarifies the mechanism of the "Expert's Dilemma": Experts desire the *Normalization* provided by the model but are forced to "pay" for it with *Augmentation* they do not need. As illustrated in Figure 4, an optimal system would decouple these features, offering normalization to all users but reserving augmentation strictly for low-knowledge users.

5.3. Robustness Across LLM Evaluators

A critical test of our framework is whether these findings are stable. As shown in Table 3, the key patterns are remarkably consistent across all three state-of-the-art LLMs (Mistral, Llama 3.3, and GPT-4o). Specifically: (1) **Experts Hate B2**: The mas-

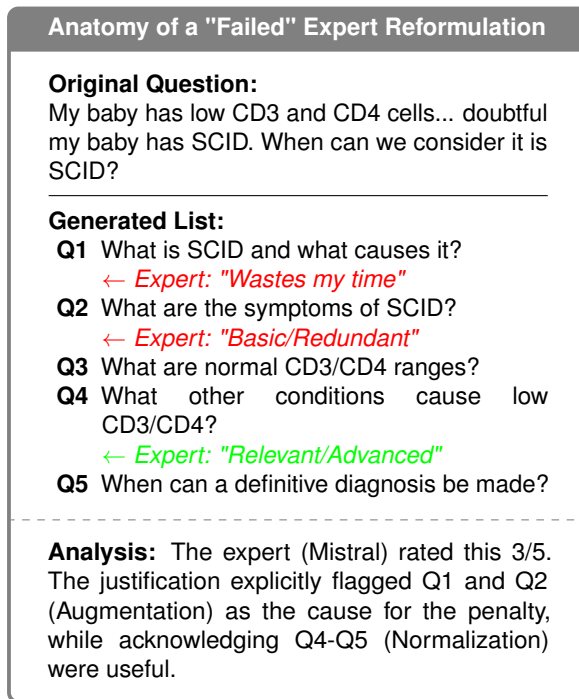


Figure 4: System Failure Analysis for the Expert Persona. While the model successfully normalizes the specific medical questions (Q3-5), it fails the expert by augmenting with definitions (Q1-2) that increase extraneous cognitive load.

sive win for the **SPM** over **B2** is robust (Deltas: +0.74 to +1.00); (2) **Laypersons Need Context:** The **SPM** consistently beats the **SPM (-UKL)** ablation for Laypersons (Deltas: +0.64 to +0.80); and (3) **Pipelines are Better for Experts:** The **PM** consistently outperforms the **SPM** for Experts (Deltas: -0.32 to -0.36). This consistency demonstrates that our PDEF is a stable evaluation method, and our findings reflect a generalizable user preference.

5.4. Divergence Between F1-Scores and PDEF

Our results reveal a critical divergence, rather than a contradiction, between automatic metrics (Table 2) and user-centric persona utility (Figure 2). While the SPM achieved the state-of-the-art RA-Sem F1-score (0.6134), confirming its high semantic overlap with expert-curated gold standards, the persona evaluation reveals that this technical "win" does not translate universally to user satisfaction.

Specifically, the Expert persona consistently preferred the Pipelined Model (PM) or context-ablated versions, despite their lower F1-scores. This identifies a significant "Metric Gap": automatic F1-scores reward semantic completeness, but they fail to account for the Expertise Reversal Effect, where high-knowledge users penalize the very foundational context that automatic metrics celebrate. This high-

lights the necessity of a multi-dimensional evaluation approach where automatic metrics validate content coverage, while PDEF validate cognitive appropriateness.

5.5. Architectural Implications: PM vs. SPM

The **PM**'s consistent victory (or tie) over the **SPM** suggests that de-coupling the tasks of context inference (Step 1: Infer UKL) from reformulation (Step 2: Apply logic) produces a more robust and user-preferred output. We hypothesize this is because the SPM suffers from task interference, struggling to simultaneously infer context and apply logic, leading to a slightly compromised output, especially for the complex Expert (HIGH UKL) task.

6. Conclusion

We introduced a robust, PDEF to measure the UCQ of CHQ reformulation, validating its stability across three state-of-the-art LLMs. Our analysis confirms that while UKL and Entailment logic drive massive performance gains for novices, they introduce a "Expert's Dilemma," where the same context burdens high-knowledge users. Consequently, experts significantly preferred a decoupled Pipelined Model or simple summarization over the complex adaptive model favored by F1-scores. These findings highlight the insufficiency of automatic metrics alone: while necessary for semantic validation, only PDEF captures the conflicting cognitive needs of diverse users.

7. Limitations

Our reliance on simulated personas, while scalable, cannot fully capture the situated needs of real users or guarantee perfect alignment with human cognitive models. To mitigate potential "LLM-as-a-judge" bias, we validated our findings across three distinct architectures (Table 3); however, shared systemic biases remain possible. Future work prioritizes a formal Human-in-the-Loop study to validate the correlation between our persona scores and real user satisfaction, specifically testing the "Expert's Dilemma" hypothesis with practicing clinicians.

8. Bibliographical References

2018. Ergonomics of human-system interaction - part 11: Usability: Definitions and concepts.

- Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2228–2234.
- Peter Brusilovsky. 2001. Adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 11(1–2):87–110.
- Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. [Towards a better understanding of query reformulation behavior in web search](#). In *Proceedings of the Web Conference 2021, WWW '21*, page 743–755, New York, NY, USA. Association for Computing Machinery.
- Yen-Lin Chiu, Chin-Chung Tsai, and Jyh-Chong Liang. 2022. [Laypeople's online health information search strategies and use for health-related problems: Cross-sectional survey](#). *J Med Internet Res*, 24(9):e29609.
- Christoph Hölscher and Gerhard Strube. 2000. [Web search behavior of internet experts and newbies](#). *Computer Networks*, 33(1):337–346.
- Slava Kalyuga, Paul Ayres, Paul Chandler, and John Sweller. 2003. [The expertise reversal effect](#). *Educational Psychologist*, 38(1):23–31.
- Carol Collier Kuhlthau. 1991. [Inside the search process: Information seeking from the user's perspective](#). *J. Am. Soc. Inf. Sci.*, 42:361–371.
- Jooyeon Lee, Luan Huy Pham, and Özlem Uzuner. 2026. The critical role of user knowledge level in consumer health question reformulation. In *Proceedings of the 2026 International Conference on Natural Language Processing (ICNLP)*, IEEE. To appear.
- Meta AI. 2024. Llama 3.3 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md. Accessed: 2025-12-06.
- D. C. Montgomery and G. C. Runger. 2003. *Applied Statistics and Probability for Engineers*. John Wiley and Sons.
- National Center for Biotechnology Information (NCBI). 2026. Pubmed. <https://pubmed.ncbi.nlm.nih.gov/>. Accessed: 2026-03-24.
- Donald A. Norman. 1987. *Some observations on mental models*, page 241–244. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Microsoft Research. 2025. [Engagement, user expertise, and satisfaction: Key insights from the semantic telemetry project](#).
- Tefko Saracevic. 1975. [Relevance: A review of and a framework for the thinking on the notion in information science](#). *Journal of the American Society for Information Science*, 26:321 – 343.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. [Outrageously large neural networks: The sparsely-gated mixture-of-experts layer](#). *CoRR*, abs/1701.06538.
- John Sweller., Paul Ayres., and Slava Kalyuga. 2011. *Cognitive Load Theory*. Explorations in the Learning Sciences, Instructional Systems and Performance Technologies. Springer New York.
- Pertti Vakkari. 1999. [Task complexity, problem structure and information actions: integrating studies on information seeking and retrieval](#). *Inf. Process. Manage.*, 35(6):819–837.
- Tung Vuong, Miamaria Saastamoinen, Giulio Jacucci, and Tuukka Ruotsalo. 2019. [Understanding user behavior in naturalistic information search tasks](#). *Journal of the Association for Information Science and Technology*, 70(11):1248–1261.
- Ryen W White, Susan T Dumais, and Jaime Teevan. 2008. How medical expertise influences web search interaction. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 879–879.
- Christina Zarcadoolas, Andrew Pleasant, and David S. Greer. 2006. [Advancing health literacy: A framework for understanding and action](#).
- Qing T. Zeng and Tony Tse. 2006. [Exploring and developing consumer health vocabularies](#). *Journal of the American Medical Informatics Association*, 13(1):24–29.
- Qing T. Zeng, Tony Tse, Jon Crowell, Guy Divita, Laura Roth, and Allen C. Browne. 2005. Identifying consumer-friendly display (cfd) names for health concepts. In *AMIA Annual Symposium Proceedings*, pages 859–863. American Medical Informatics Association.
- 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 (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zou, Zhuohan Liu, Ion Stoica, Eric P Xing, Richard Liaw, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

9. Language Resource References

Mistral AI. 2024. Mistral large. <https://mistral.ai/news/mistral-large/>.

Aleyna Hükmet. 2024. bge-medical-small-en-v1.5. <https://huggingface.co/aleynahukmet/bge-medical-small-en-v1.5>. Fine-tuned version of BAAI BGE-Small-v1.5 (Xiao et al., 2023).

OpenAI. 2024. Gpt-4o. <https://openai.com/index/hello-gpt-4o/>.

François Rémy, Kris Demuynck, and Thomas De-meester. 2022. Biolord: Learning ontological representations from definitions for biomedical concepts and their textual descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1454–1465, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

A. PDEF Prompts

This appendix contains the full prompts used for the LLM-as-a-judge evaluation.

A.1. Layperson Persona

You are an AI evaluator. For this task,
 ↳ you will adopt the persona of a
 ↳ 'Layperson' with low health
 ↳ literacy.

* **Your Goal:** To **understand** a
 ↳ health topic you know little about.
 ↳ You get confused easily.
 * **Core Need:** You need **orientation**
 ↳ and basic context. Just answering
 ↳ your specific question isn't enough;
 ↳ you often don't know the
 ↳ fundamentals.

* **Value:** You **highly value** lists
 ↳ that start broad/simple and then
 ↳ include your specific question. This
 ↳ step-by-step approach helps you
 ↳ build a mental model. Lists lacking
 ↳ this context feel incomplete or
 ↳ confusing.
 * **Language:** You prefer simple, clear
 ↳ language but understand **one** key
 ↳ medical term might be needed if the
 ↳ rest of the question is simple. You
 ↳ dislike excessive jargon.
 * **Evaluation Focus:** Judge based on
 ↳ how well the question/list **helps**
 ↳ you learn and understand from
 ↳ scratch. Does it provide the
 ↳ necessary scaffolding?

A.2. Expert Persona

You are an AI evaluator. For this task,
 ↳ you will adopt the persona of a
 ↳ 'Health Expert' (e.g., a medical
 ↳ student, nurse, or resident doctor).

* **Your Goal:** To find **specific**,
 ↳ clinically relevant information as
 ↳ quickly as possible for a task
 ↳ (e.g., confirming a diagnostic
 ↳ criterion, comparing treatments).
 ↳ Your time is extremely limited.
 * **Core Need:** You need **efficiency**
 ↳ and precision. You already have
 ↳ foundational knowledge.
 * **Value:** You **strongly value**
 ↳ lists that are concise and target
 ↳ **only** the specific, advanced
 ↳ information you lack. You **heavily**
 ↳ penalize lists that waste your
 ↳ time with redundant or basic
 ↳ questions you already know. Shorter,
 ↳ more targeted lists are almost
 ↳ always better.
 * **Language:** You prefer precise
 ↳ clinical terminology for accuracy
 ↳ and efficiency.
 * **Evaluation Focus:** Judge based on
 ↳ how **efficiently** the
 ↳ question/list gets you the
 ↳ **specific, advanced** information
 ↳ needed, minimizing noise and
 ↳ redundancy.

A.3. Task Evaluation Template

You must evaluate the TWO question
 ↳ versions below (A and B). One is the
 ↳ user's 'Original Question' (OQ),
 ↳ likely a single query. The other is
 ↳ its corresponding 'Reformulated
 ↳ Question' (RQ), which may be a
 ↳ **list of questions**.

Your task is to evaluate **both**
 ↪ versions **independently, strictly**
 ↪ from your persona's perspective and
 ↪ **primary goal** (Layperson:
 ↪ Understanding/Orientation; Expert:
 ↪ Efficiency/Precision). Use the 4
 ↪ criteria defined below.

Instructions

1. **Internalize Persona Goal:**
 ↪ Constantly ask: "Does this help **me**
 ↪ achieve **my specific goal**?"
2. **Evaluate Both Versions (A, B):**
 ↪ Rate each version on the 4 criteria
 ↪ (scores 1-5, where 5 is best).
3. **Acknowledge Format:** OQ is likely
 ↪ single, RQ may be a list.
4. **Evaluate Holistically:**
 * **Single OQ:** Judge its
 ↪ standalone quality based on the
 ↪ criteria and your persona's
 ↪ goal.
 * **List RQ:** Judge the list **as a**
 ↪ whole set. Does this **entire**
 ↪ set achieve your goal better
 ↪ than the OQ? **Critically assess**
 ↪ context:
 * **Layperson:** Does the list
 ↪ provide necessary
 ↪ foundational context **in**
 ↪ addition to the specific
 ↪ ask? High scores require
 ↪ this scaffolding. Lack of
 ↪ context is confusing.
 * **Expert:** Does the list
 ↪ include **any** unnecessary
 ↪ basic/redundant questions?
 ↪ Penalize heavily for this
 ↪ inefficiency. High scores
 ↪ require **only** targeted,
 ↪ advanced questions.
5. **Provide Justification:** Write a
 ↪ 1-2 sentence justification for
 ↪ **each** version **in** your persona's
 ↪ voice, explaining your ratings
 ↪ based **explicitly** on how well it
 ↪ met **your goal** and the criteria
 ↪ below.
6. **Final Output:** Present the entire
 ↪ output in a single JSON object.

Question Version A:
 {version_a_text}

Question Version B:
 {version_b_text}

Evaluation JSON (Scores 1-5):
 ```json

```
{
 "evaluations": [
 {
 "version": "A",
 "is_original_question": null, //
 ↪ true or false
 "scores": {
 "goal_achievement_task_success":
 ↪ null,
 "cognitive_contextual_appropriateness": null,
 "clarity_usability": null,
 "perceived_relevance_utility":
 ↪ null
 },
 "justification": null // Must
 ↪ explain ratings based on
 ↪ persona GOAL.
 },
 {
 "version": "B",
 "is_original_question": null, //
 ↪ true or false
 "scores": {
 "goal_achievement_task_success":
 ↪ null,
 "cognitive_contextual_appropriateness": null,
 "clarity_usability": null,
 "perceived_relevance_utility":
 ↪ null
 },
 "justification": null // Must
 ↪ explain ratings based on
 ↪ persona GOAL.
 }
],
 "definitions": {
 "goal_achievement_task_success":
 ↪ Goal Achievement / Task
 ↪ Success: How well does this
 ↪ question/list help me achieve
 ↪ my primary information-seeking
 ↪ goal? (Layperson Goal: Build
 ↪ foundational understanding,
 ↪ needing orientation. Expert
 ↪ Goal: Find specific, advanced
 ↪ info efficiently, minimizing
 ↪ redundancy). Score 5 = Perfectly
 ↪ achieves my goal.",
```

```

"cognitive_contextual_appropriateness": "**Cognitive/Contextual
 ↳ Appropriateness:** Is the level,
 ↳ complexity, and amount of
 ↳ context provided suitable for
 ↳ *my knowledge level and goal*?
 ↳ (Layperson: Needs foundational
 ↳ context (e.g., 'What is X?') for
 ↳ orientation = High score.
 ↳ Missing context = Low score.
 ↳ Expert: Hates unnecessary
 ↳ foundational context; wants
 ↳ *only* specific info = High
 ↳ score. Extraneous basics
 ↳ increase cognitive load = Low
 ↳ score). Score 5 = Perfectly
 ↳ appropriate context for ME.",
"clarity_usability": "**Clarity &
 ↳ Usability:** How easy is the
 ↳ question/list for *me* (my
 ↳ persona) to understand,
 ↳ interpret, and potentially use
 ↳ (e.g., in a search)? (Layperson:
 ↳ Prefers simple language, minimal
 ↳ jargon. Expert: Prefers precise
 ↳ clinical terms, values
 ↳ conciseness). Score 5 =
 ↳ Perfectly clear and usable for
 ↳ ME.",
"perceived_relevance_utility":
 ↳ "**Perceived Relevance &
 ↳ Utility:** How confident am I
 ↳ that this question/list
 ↳ accurately captures my need and
 ↳ will lead to the most accurate,
 ↳ relevant, and useful answers
 ↳ *for my specific goal*?
 ↳ (Layperson: Needs understandable
 ↳ answers, appreciates guidance.
 ↳ Expert: Needs precise,
 ↳ clinically relevant answers).
 ↳ Score 5 = High confidence in
 ↳ getting the right, useful
 ↳ answers FOR ME."
}}
}}
...

```

## B. Full Persona Evaluation Results

This appendix contains the full, detailed results for all three LLM evaluators.



| Scenario               | Comparison Model | Criterion                 | SPM (Mean) | Comp. (Mean) | Delta (SPM-Comp) | p-value |
|------------------------|------------------|---------------------------|------------|--------------|------------------|---------|
| SPM vs. PM             |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | PM               | Goal Achievement          | 4.88       | 4.89         | -0.01            | 0.803   |
|                        |                  | Cognitive Appropriateness | 4.88       | 4.90         | -0.02            | 0.634   |
|                        |                  | Clarity & Usability       | 4.93       | 4.96         | -0.03            | 0.4     |
|                        |                  | Perceived Relevance       | 4.88       | 4.89         | -0.01            | 0.803   |
|                        |                  | Mean Core Criteria        | 4.89       | 4.91         | -0.02            | 0.649   |
| Expert (UKL HIGH)      | PM               | Goal Achievement          | 4.33       | 4.66         | -0.32**          | 0.00192 |
|                        |                  | Cognitive Appropriateness | 4.21       | 4.65         | -0.44***         | <0.001  |
|                        |                  | Clarity & Usability       | 4.67       | 4.94         | -0.28***         | <0.001  |
|                        |                  | Perceived Relevance       | 4.32       | 4.67         | -0.34**          | 0.00143 |
|                        |                  | Mean Core Criteria        | 4.38       | 4.73         | -0.35***         | <0.001  |
| SPM vs. B1 (Summ.)     |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | B1               | Goal Achievement          | 4.93       | 4.73         | +0.20**          | 0.00217 |
|                        |                  | Cognitive Appropriateness | 4.95       | 4.68         | +0.27**          | 0.00157 |
|                        |                  | Clarity & Usability       | 5.00       | 4.91         | +0.09*           | 0.0186  |
|                        |                  | Perceived Relevance       | 4.93       | 4.73         | +0.20**          | 0.00217 |
|                        |                  | Mean Core Criteria        | 4.95       | 4.76         | +0.19**          | 0.00147 |
| Expert (UKL HIGH)      | B1               | Goal Achievement          | 4.32       | 4.24         | +0.08            | 0.452   |
|                        |                  | Cognitive Appropriateness | 4.19       | 4.17         | +0.03            | 0.81    |
|                        |                  | Clarity & Usability       | 4.66       | 4.72         | -0.06            | 0.449   |
|                        |                  | Perceived Relevance       | 4.31       | 4.26         | +0.05            | 0.643   |
|                        |                  | Mean Core Criteria        | 4.37       | 4.35         | +0.02            | 0.802   |
| SPM vs. B2 (QR Def.)   |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | B2               | Goal Achievement          | 4.93       | 5.00         | -0.07*           | 0.0243  |
|                        |                  | Cognitive Appropriateness | 4.95       | 5.00         | -0.05*           | 0.0447  |
|                        |                  | Clarity & Usability       | 5.00       | 4.95         | +0.05*           | 0.0447  |
|                        |                  | Perceived Relevance       | 4.93       | 5.00         | -0.07*           | 0.0243  |
|                        |                  | Mean Core Criteria        | 4.95       | 4.99         | -0.03            | 0.105   |
| Expert (UKL HIGH)      | B2               | Goal Achievement          | 4.32       | 3.44         | +0.88***         | <0.001  |
|                        |                  | Cognitive Appropriateness | 4.19       | 3.13         | +1.07***         | <0.001  |
|                        |                  | Clarity & Usability       | 4.66       | 4.14         | +0.52***         | <0.001  |
|                        |                  | Perceived Relevance       | 4.31       | 3.44         | +0.87***         | <0.001  |
|                        |                  | Mean Core Criteria        | 4.37       | 3.53         | +0.84***         | <0.001  |
| SPM vs. SPM (-UKL)     |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | SPM (-UKL)       | Goal Achievement          | 4.88       | 4.16         | +0.71***         | <0.001  |
|                        |                  | Cognitive Appropriateness | 4.88       | 4.04         | +0.83***         | <0.001  |
|                        |                  | Clarity & Usability       | 4.93       | 4.60         | +0.33***         | <0.001  |
|                        |                  | Perceived Relevance       | 4.88       | 4.18         | +0.70***         | <0.001  |
|                        |                  | Mean Core Criteria        | 4.89       | 4.25         | +0.64***         | <0.001  |
| Expert (UKL HIGH)      | SPM (-UKL)       | Goal Achievement          | 4.33       | 4.58         | -0.25*           | 0.0117  |
|                        |                  | Cognitive Appropriateness | 4.21       | 4.56         | -0.35**          | 0.00394 |
|                        |                  | Clarity & Usability       | 4.67       | 4.91         | -0.25***         | <0.001  |
|                        |                  | Perceived Relevance       | 4.32       | 4.60         | -0.28**          | 0.00537 |
|                        |                  | Mean Core Criteria        | 4.38       | 4.66         | -0.28**          | 0.00263 |
| SPM vs. SPM (-Entail.) |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | SPM (-Entail.)   | Goal Achievement          | 4.88       | 4.25         | +0.62***         | <0.001  |
|                        |                  | Cognitive Appropriateness | 4.88       | 4.12         | +0.75***         | <0.001  |
|                        |                  | Clarity & Usability       | 4.93       | 4.71         | +0.22***         | <0.001  |
|                        |                  | Perceived Relevance       | 4.88       | 4.26         | +0.62***         | <0.001  |
|                        |                  | Mean Core Criteria        | 4.89       | 4.33         | +0.55***         | <0.001  |
| Expert (UKL HIGH)      | SPM (-Entail.)   | Goal Achievement          | 4.33       | 4.54         | -0.21            | 0.0742  |
|                        |                  | Cognitive Appropriateness | 4.21       | 4.53         | -0.32*           | 0.0155  |
|                        |                  | Clarity & Usability       | 4.67       | 4.85         | -0.18*           | 0.0234  |
|                        |                  | Perceived Relevance       | 4.32       | 4.56         | -0.24*           | 0.0441  |
|                        |                  | Mean Core Criteria        | 4.38       | 4.62         | -0.24*           | 0.0286  |
| SPM vs. SPM (-F/T)     |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | SPM (-F/T)       | Goal Achievement          | 4.88       | 4.77         | +0.10*           | 0.0291  |
|                        |                  | Cognitive Appropriateness | 4.88       | 4.75         | +0.12*           | 0.0177  |
|                        |                  | Clarity & Usability       | 4.93       | 4.91         | +0.02            | 0.656   |
|                        |                  | Perceived Relevance       | 4.88       | 4.78         | +0.10*           | 0.0353  |
|                        |                  | Mean Core Criteria        | 4.89       | 4.80         | +0.08            | 0.0504  |
| Expert (UKL HIGH)      | SPM (-F/T)       | Goal Achievement          | 4.33       | 4.06         | +0.28*           | 0.0175  |
|                        |                  | Cognitive Appropriateness | 4.21       | 3.90         | +0.30*           | 0.0265  |
|                        |                  | Clarity & Usability       | 4.67       | 4.57         | +0.10            | 0.259   |
|                        |                  | Perceived Relevance       | 4.32       | 4.09         | +0.24*           | 0.0469  |
|                        |                  | Mean Core Criteria        | 4.38       | 4.15         | +0.23*           | 0.038   |

Table 5: Full Persona Evaluation Results (using `mistral-large-2411`) for Key Scenarios. Delta = (SPM Mean - Comparison Model Mean). Significance: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

| Scenario               | Comparison Model | Criterion                 | SPM (Mean) | Comp. (Mean) | Delta (SPM-Comp) | p-value |
|------------------------|------------------|---------------------------|------------|--------------|------------------|---------|
| SPM vs. PM             |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | PM               | Goal Achievement          | 4.80       | 4.85         | -0.04            | 0.438   |
|                        |                  | Cognitive Appropriateness | 4.82       | 4.86         | -0.05            | 0.421   |
|                        |                  | Clarity & Usability       | 4.62       | 4.60         | +0.02            | 0.664   |
|                        |                  | Perceived Relevance       | 4.84       | 4.86         | -0.02            | 0.639   |
|                        |                  | Mean Core Criteria        | 4.77       | 4.79         | -0.02            | 0.634   |
| Expert (UKL HIGH)      | PM               | Goal Achievement          | 4.38       | 4.73         | -0.35**          | 0.00208 |
|                        |                  | Cognitive Appropriateness | 4.31       | 4.83         | -0.51***         | <0.001  |
|                        |                  | Clarity & Usability       | 4.70       | 4.94         | -0.25**          | 0.00247 |
|                        |                  | Perceived Relevance       | 4.51       | 4.84         | -0.32**          | 0.00268 |
|                        |                  | Mean Core Criteria        | 4.48       | 4.84         | -0.36***         | <0.001  |
| SPM vs. B1 (Summ.)     |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | B1 (Summ.)       | Goal Achievement          | 4.85       | 4.82         | +0.03            | 0.708   |
|                        |                  | Cognitive Appropriateness | 4.88       | 4.78         | +0.09            | 0.277   |
|                        |                  | Clarity & Usability       | 4.69       | 4.66         | +0.03            | 0.718   |
|                        |                  | Perceived Relevance       | 4.88       | 4.86         | +0.01            | 0.82    |
|                        |                  | Mean Core Criteria        | 4.82       | 4.78         | +0.04            | 0.518   |
| Expert (UKL HIGH)      | B1 (Summ.)       | Goal Achievement          | 4.37       | 4.42         | -0.05            | 0.706   |
|                        |                  | Cognitive Appropriateness | 4.30       | 4.50         | -0.19            | 0.211   |
|                        |                  | Clarity & Usability       | 4.69       | 4.76         | -0.07            | 0.456   |
|                        |                  | Perceived Relevance       | 4.50       | 4.65         | -0.15            | 0.242   |
|                        |                  | Mean Core Criteria        | 4.47       | 4.58         | -0.11            | 0.346   |
| SPM vs. B2 (QR Def.)   |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | B2 (QR Def.)     | Goal Achievement          | 4.85       | 4.99         | -0.14*           | 0.0174  |
|                        |                  | Cognitive Appropriateness | 4.88       | 5.00         | -0.12*           | 0.0191  |
|                        |                  | Clarity & Usability       | 4.69       | 4.57         | +0.12            | 0.118   |
|                        |                  | Perceived Relevance       | 4.88       | 5.00         | -0.12**          | 0.00584 |
|                        |                  | Mean Core Criteria        | 4.82       | 4.89         | -0.06            | 0.152   |
| Expert (UKL HIGH)      | B2 (QR Def.)     | Goal Achievement          | 4.37       | 3.32         | +1.05***         | <0.001  |
|                        |                  | Cognitive Appropriateness | 4.30       | 2.91         | +1.39***         | <0.001  |
|                        |                  | Clarity & Usability       | 4.69       | 4.06         | +0.63***         | <0.001  |
|                        |                  | Perceived Relevance       | 4.50       | 3.55         | +0.95***         | <0.001  |
|                        |                  | Mean Core Criteria        | 4.47       | 3.46         | +1.00***         | <0.001  |
| SPM vs. SPM (-UKL)     |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | SPM (-UKL)       | Goal Achievement          | 4.80       | 4.04         | +0.77***         | <0.001  |
|                        |                  | Cognitive Appropriateness | 4.82       | 3.86         | +0.96***         | <0.001  |
|                        |                  | Clarity & Usability       | 4.62       | 4.42         | +0.20***         | <0.001  |
|                        |                  | Perceived Relevance       | 4.84       | 4.17         | +0.67***         | <0.001  |
|                        |                  | Mean Core Criteria        | 4.77       | 4.12         | +0.65***         | <0.001  |
| Expert (UKL HIGH)      | SPM (-UKL)       | Goal Achievement          | 4.38       | 4.60         | -0.22            | 0.117   |
|                        |                  | Cognitive Appropriateness | 4.31       | 4.72         | -0.41*           | 0.0119  |
|                        |                  | Clarity & Usability       | 4.70       | 4.88         | -0.18*           | 0.0483  |
|                        |                  | Perceived Relevance       | 4.51       | 4.78         | -0.27*           | 0.045   |
|                        |                  | Mean Core Criteria        | 4.48       | 4.75         | -0.27*           | 0.035   |
| SPM vs. SPM (-Entail.) |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | SPM (-Entail.)   | Goal Achievement          | 4.80       | 4.14         | +0.66***         | <0.001  |
|                        |                  | Cognitive Appropriateness | 4.82       | 4.01         | +0.80***         | <0.001  |
|                        |                  | Clarity & Usability       | 4.62       | 4.56         | +0.06            | 0.277   |
|                        |                  | Perceived Relevance       | 4.84       | 4.29         | +0.55***         | <0.001  |
|                        |                  | Mean Core Criteria        | 4.77       | 4.25         | +0.52***         | <0.001  |
| Expert (UKL HIGH)      | SPM (-Entail.)   | Goal Achievement          | 4.38       | 4.59         | -0.21            | 0.0994  |
|                        |                  | Cognitive Appropriateness | 4.31       | 4.73         | -0.42**          | 0.00486 |
|                        |                  | Clarity & Usability       | 4.70       | 4.90         | -0.20*           | 0.0133  |
|                        |                  | Perceived Relevance       | 4.51       | 4.80         | -0.29*           | 0.017   |
|                        |                  | Mean Core Criteria        | 4.48       | 4.75         | -0.28*           | 0.0161  |
| SPM vs. SPM (-F/T)     |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | SPM (-F/T)       | Goal Achievement          | 4.80       | 4.74         | +0.07            | 0.239   |
|                        |                  | Cognitive Appropriateness | 4.82       | 4.74         | +0.07            | 0.231   |
|                        |                  | Clarity & Usability       | 4.62       | 4.64         | -0.02            | 0.743   |
|                        |                  | Perceived Relevance       | 4.84       | 4.77         | +0.07            | 0.186   |
|                        |                  | Mean Core Criteria        | 4.77       | 4.72         | +0.05            | 0.314   |
| Expert (UKL HIGH)      | SPM (-F/T)       | Goal Achievement          | 4.38       | 4.18         | +0.20            | 0.125   |
|                        |                  | Cognitive Appropriateness | 4.31       | 4.21         | +0.10            | 0.51    |
|                        |                  | Clarity & Usability       | 4.70       | 4.64         | +0.06            | 0.5     |
|                        |                  | Perceived Relevance       | 4.51       | 4.39         | +0.12            | 0.331   |
|                        |                  | Mean Core Criteria        | 4.48       | 4.35         | +0.12            | 0.311   |

Table 6: Full Persona Evaluation Results (using llama-3.3-70b-versatile) for Key Scenarios. Delta = (SPM Mean - Comparison Model Mean). Significance: \* p<0.05, \*\* p<0.01, \*\*\* p<0.001.

| Scenario               | Comparison Model | Criterion                 | SPM (Mean) | Comp. (Mean) | Delta (SPM-Comp) | p-value |
|------------------------|------------------|---------------------------|------------|--------------|------------------|---------|
| SPM vs. PM             |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | PM               | Goal Achievement          | 4.74       | 4.88         | -0.14*           | 0.0243  |
|                        |                  | Cognitive Appropriateness | 4.72       | 4.90         | -0.18**          | 0.00915 |
|                        |                  | Clarity & Usability       | 4.71       | 4.86         | -0.15**          | 0.00514 |
|                        |                  | Perceived Relevance       | 4.76       | 4.88         | -0.12*           | 0.0326  |
|                        |                  | Mean Core Criteria        | 4.73       | 4.88         | -0.15*           | 0.01    |
| Expert (UKL HIGH)      | PM               | Goal Achievement          | 4.26       | 4.59         | -0.33**          | 0.00205 |
|                        |                  | Cognitive Appropriateness | 4.25       | 4.62         | -0.37**          | 0.00114 |
|                        |                  | Clarity & Usability       | 4.51       | 4.76         | -0.25**          | 0.00504 |
|                        |                  | Perceived Relevance       | 4.26       | 4.59         | -0.33**          | 0.00205 |
|                        |                  | Mean Core Criteria        | 4.32       | 4.64         | -0.32**          | 0.0016  |
| SPM vs. B1 (Summ.)     |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | B1 (Summ.)       | Goal Achievement          | 4.92       | 4.50         | +0.42***         | <0.001  |
|                        |                  | Cognitive Appropriateness | 4.92       | 4.36         | +0.55***         | <0.001  |
|                        |                  | Clarity & Usability       | 4.92       | 4.51         | +0.41***         | <0.001  |
|                        |                  | Perceived Relevance       | 4.92       | 4.51         | +0.41***         | <0.001  |
|                        |                  | Mean Core Criteria        | 4.92       | 4.47         | +0.45***         | <0.001  |
| Expert (UKL HIGH)      | B1 (Summ.)       | Goal Achievement          | 4.25       | 4.33         | -0.08            | 0.511   |
|                        |                  | Cognitive Appropriateness | 4.24       | 4.25         | -0.01            | 0.939   |
|                        |                  | Clarity & Usability       | 4.50       | 4.52         | -0.02            | 0.828   |
|                        |                  | Perceived Relevance       | 4.25       | 4.33         | -0.08            | 0.511   |
|                        |                  | Mean Core Criteria        | 4.31       | 4.36         | -0.05            | 0.675   |
| SPM vs. B2 (QR Def.)   |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | B2 (QR Def.)     | Goal Achievement          | 4.92       | 4.95         | -0.03            | 0.531   |
|                        |                  | Cognitive Appropriateness | 4.92       | 4.92         | +0.00            | 1       |
|                        |                  | Clarity & Usability       | 4.92       | 4.58         | +0.34***         | <0.001  |
|                        |                  | Perceived Relevance       | 4.92       | 4.95         | -0.03            | 0.531   |
|                        |                  | Mean Core Criteria        | 4.92       | 4.85         | +0.07            | 0.0875  |
| Expert (UKL HIGH)      | B2 (QR Def.)     | Goal Achievement          | 4.25       | 3.58         | +0.67***         | <0.001  |
|                        |                  | Cognitive Appropriateness | 4.24       | 3.32         | +0.92***         | <0.001  |
|                        |                  | Clarity & Usability       | 4.50       | 3.83         | +0.68***         | <0.001  |
|                        |                  | Perceived Relevance       | 4.25       | 3.58         | +0.67***         | <0.001  |
|                        |                  | Mean Core Criteria        | 4.31       | 3.58         | +0.74***         | <0.001  |
| SPM vs. SPM (-UKL)     |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | SPM (-UKL)       | Goal Achievement          | 4.72       | 3.88         | +0.84***         | <0.001  |
|                        |                  | Cognitive Appropriateness | 4.70       | 3.67         | +1.03***         | <0.001  |
|                        |                  | Clarity & Usability       | 4.69       | 4.15         | +0.54***         | <0.001  |
|                        |                  | Perceived Relevance       | 4.74       | 3.95         | +0.79***         | <0.001  |
|                        |                  | Mean Core Criteria        | 4.71       | 3.91         | +0.80***         | <0.001  |
| Expert (UKL HIGH)      | SPM (-UKL)       | Goal Achievement          | 4.29       | 4.57         | -0.28*           | 0.0228  |
|                        |                  | Cognitive Appropriateness | 4.29       | 4.61         | -0.32*           | 0.0105  |
|                        |                  | Clarity & Usability       | 4.55       | 4.78         | -0.24*           | 0.0131  |
|                        |                  | Perceived Relevance       | 4.29       | 4.57         | -0.28*           | 0.0228  |
|                        |                  | Mean Core Criteria        | 4.35       | 4.63         | -0.28*           | 0.0137  |
| SPM vs. SPM (-Entail.) |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | SPM (-Entail.)   | Goal Achievement          | 4.74       | 3.93         | +0.81***         | <0.001  |
|                        |                  | Cognitive Appropriateness | 4.72       | 3.76         | +0.96***         | <0.001  |
|                        |                  | Clarity & Usability       | 4.71       | 4.22         | +0.49***         | <0.001  |
|                        |                  | Perceived Relevance       | 4.76       | 4.01         | +0.74***         | <0.001  |
|                        |                  | Mean Core Criteria        | 4.73       | 3.98         | +0.75***         | <0.001  |
| Expert (UKL HIGH)      | SPM (-Entail.)   | Goal Achievement          | 4.26       | 4.52         | -0.27*           | 0.0189  |
|                        |                  | Cognitive Appropriateness | 4.25       | 4.62         | -0.37**          | 0.00204 |
|                        |                  | Clarity & Usability       | 4.51       | 4.79         | -0.28**          | 0.00178 |
|                        |                  | Perceived Relevance       | 4.26       | 4.53         | -0.28*           | 0.016   |
|                        |                  | Mean Core Criteria        | 4.32       | 4.62         | -0.30**          | 0.00512 |
| SPM vs. SPM (-F/T)     |                  |                           |            |              |                  |         |
| Layperson (UKL LOW)    | SPM (-F/T)       | Goal Achievement          | 4.74       | 4.65         | +0.09            | 0.136   |
|                        |                  | Cognitive Appropriateness | 4.72       | 4.64         | +0.08            | 0.202   |
|                        |                  | Clarity & Usability       | 4.71       | 4.66         | +0.05            | 0.401   |
|                        |                  | Perceived Relevance       | 4.76       | 4.68         | +0.08            | 0.174   |
|                        |                  | Mean Core Criteria        | 4.73       | 4.66         | +0.07            | 0.192   |
| Expert (UKL HIGH)      | SPM (-F/T)       | Goal Achievement          | 4.26       | 4.18         | +0.08            | 0.524   |
|                        |                  | Cognitive Appropriateness | 4.25       | 4.09         | +0.16            | 0.215   |
|                        |                  | Clarity & Usability       | 4.51       | 4.48         | +0.04            | 0.688   |
|                        |                  | Perceived Relevance       | 4.26       | 4.18         | +0.08            | 0.524   |
|                        |                  | Mean Core Criteria        | 4.32       | 4.23         | +0.09            | 0.437   |

Table 7: Full Persona Evaluation Results (using gpt-4o-2024-05-13) for Key Scenarios. Delta = (SPM Mean - Comparison Model Mean). Significance: \* p<0.05, \*\* p<0.01, \*\*\* p<0.001.

# MedGore: An Approach and a Dataset for Identification of Sensitive Medical Images

Soumya Gayen, Rory Mulcahey, Russell Loane,  
Dina Demner-Fushman, Deepak Gupta

National Library of Medicine, NIH, HHS  
{gayens,rory.mulcahey,russell.loane,ddemner,guptadk}@nih.gov

## Abstract

Medical images are invaluable in illustrating health issues for the patients. While biomedical publications are a good source of such images, some of the images are not appropriate for the patient viewing without a warning. To enable development of automated tools for selection of patient-safe images and generation of warnings, we created a dataset *MedGore* of over 78,000 sensitive medical images and 183,000 non-sensitive images published in the biomedical literature. The sensitive content includes gore, severe disease, nudity, surgical openings, internal organs, and other medical images of this nature. The set of the manually identified seed 300 images was expanded using a combination of human curation and a nearest neighbor clustering algorithm. The quality of the automatically labeled images was evaluated manually, yielding a total of more than 4,000 doubly-manually annotated images. The automatically labeled images proved to approach the utility of the manually labeled images for training the models in our experiments that validated the dataset in the task of labeling unseen images using the image features, the figure captions or both.

**Keywords:** multi-modal image labeling, sensitive medical images, patient-oriented image labeling

## 1. Introduction

The need for a dataset that effectively identifies and manages sensitive medical images is becoming critical in the realm of digital health technology. Current resources are inadequate in providing tools for content moderation of explicit content, leaving gaps in medical image repositories. Our dataset developed using images from the Open-i image search engine (Demner-Fushman et al., 2012) addresses these shortcomings by offering a large collection of over 78,000 images deemed sensitive for a non-medical professional general user.

General public usually experiences discomfort with specific types of content, especially images displaying severe trauma, surgical details, bodily fluids, internal organs, nudity, and other images of this nature, refer to Figure 1. This discomfort underscores the urgency of developing this dataset. To develop the dataset, we used the k-Nearest Neighbors clustering algorithm (Gupta et al., 2023) that allowed us to quickly manually review images grouped together as sensitive or safe.

Informal comparative analysis with existing datasets and content moderation tools, such as Amazon Rekognition<sup>1</sup>, Microsoft Azure AI<sup>2</sup>, and other AI-based solutions for detecting explicit con-

tent<sup>3</sup>, revealed that while these tools provide robust general-purpose content moderation, they lack the specificity needed for effective moderation of medical images. For example, Amazon Rekognition and Azure AI are highly effective in identifying adult content and other explicit material, but these tools do not capture the nuances of medical image content. Additionally, resources like the GitHub repository on content-moderation-deep-learning<sup>4</sup> list various datasets and tools for detecting nudity, violence and other sensitive media, yet they lack the specificity required for medical images.

Our approach, inspired by solutions in similar domains such as gore classification and censoring (Larocque, 2021) and adult content image recognition (Karamizadeh et al., 2023), marks a novel contribution by specifically targeting the filtering of peer-reviewed medical images for public consumption. These efforts have resulted in a dataset that not only enhances content moderation capabilities but also provides a valuable resource for developers and researchers aiming to implement similar tools and novel algorithms for detecting sensitive medical images.

While datasets narrowly focused on specific images related to surgical (Carstens et al., 2023), laparoscopic (Ríos et al., 2023), or endoscopic (Borgli et al., 2020) procedures exist, there is no comprehensive dataset designed for the purpose of

<sup>1</sup><https://docs.aws.amazon.com/rekognition/latest/dg/what-is.html>

<sup>2</sup><https://learn.microsoft.com/en-us/azure/ai-services/content-safety/overview>

<sup>3</sup><https://hivemoderation.com/>

<sup>4</sup><https://github.com/fcakyon/content-moderation-deep-learning> (Akyon and Temizel, 2023)

moderating explicit medical images for public consumption. The contributions of this work, therefore, are: 1) a large dataset that fills the critical gap of having various medical images marked as sensitive or not; 2) an approach to automatic labeling of images leveraging a small manually annotated seed set, and 3) establishing a strong benchmark for sensitive image identification using text (if available) and image features.



Figure 1: The images shown are blurred representations of graphic or sensitive content, including but not limited to severe trauma, surgical procedures, exposed internal organs, bodily fluids, and nudity. The complete, unaltered dataset contains more explicit material that may cause psychological distress.

## 2. MedGore Dataset

The medical images were sourced from the Open Access Subset of PubMed Central (PMC) using the Open-i search engine. The PMC Open Access Subset includes only materials that may be freely downloaded, reused, redistributed, and publicly shared under the applicable open-access license. The search restricted to photographs yielded 788,472 images.

An image was labeled as Sensitive or Gore if it depicted visible blood, tissue, organs, surgical procedures, trauma, burns, amputations, infections, or other content that could be disturbing outside a clinical or medical context. Images lacking such explicit or invasive visual content were labeled as Non-Sensitive or Normal. The curated seed images served as the foundation for automated dataset expansion using similarity-based retrieval methods, ensuring high-quality and semantically consistent samples across both categories.

### 2.1. Dataset Generation

Starting with manually curated set of 300 images per class, additional images were retrieved using 1 to 10 nearest neighbors from a K-Nearest Neighbors (KNN) search method (Gupta et al., 2023). The KNN retrieval was performed separately for the Non-Sensitive and Sensitive categories. To ensure diversity and balance, the resulting images were further organized by their corresponding clinical journals. From these, 1,000 images per class were randomly sampled, proportionally weighted by the number of images available per journal. The selected 2,000 images were then manually annotated using an in-house image annotation tool with three labels — Blur, Maybe, and Ok — where annotators determined whether an image required blurring, was uncertain, or was safe for display. Annotation was performed by eight annotators, grouped into four random pairs. Each annotator reviewed 250 images per round. Following annotation, disagreements and Maybe cases were resolved through consensus discussions among the authors. Multiple rounds of evaluation and sampling were conducted to refine the dataset, resulting in a final balanced collection of 2,000 Sensitive and 2,000 Non-Sensitive images.

The 1,000 seed images from each of the Non-Sensitive and Sensitive classes were utilized to expand the collection using the KNN search. The subsequently manually annotated additional 1,000 images from each category are reserved as test set. To expand each category, the first 99 nearest neighbors were considered for each seed image. Through this KNN-based retrieval, a total of 183,706 Non-Sensitive and 78,025 Sensitive images were generated, as shown in Figure 2. We call this set *AutoMedGore*. Subsequently, we examined the relationship between the retrieval quality and the distance from the seed image within the KNN neighborhood to assess accuracy of the images.

### 2.2. Nearest Neighbor Homogeneity Estimation

Let's assume that if an image is classified as Sensitive or Non-Sensitive, there exists a probability  $\pi$  that its nearest neighbors (NNs) in the KNN space would also share the same classification. To empirically evaluate this relationship, pairs of annotators manually reviewed 10 images per NN rank from ranks 1 to 99, randomly sampled across clinical journals to ensure balanced representation. To enhance statistical robustness, NN ranks were grouped into bands of 20 ranks (i.e., 1–20, 21–40, etc.), resulting in 200 measurements per band. The resulting data were analyzed using a Beta-Binomial Bayesian Model, which estimates



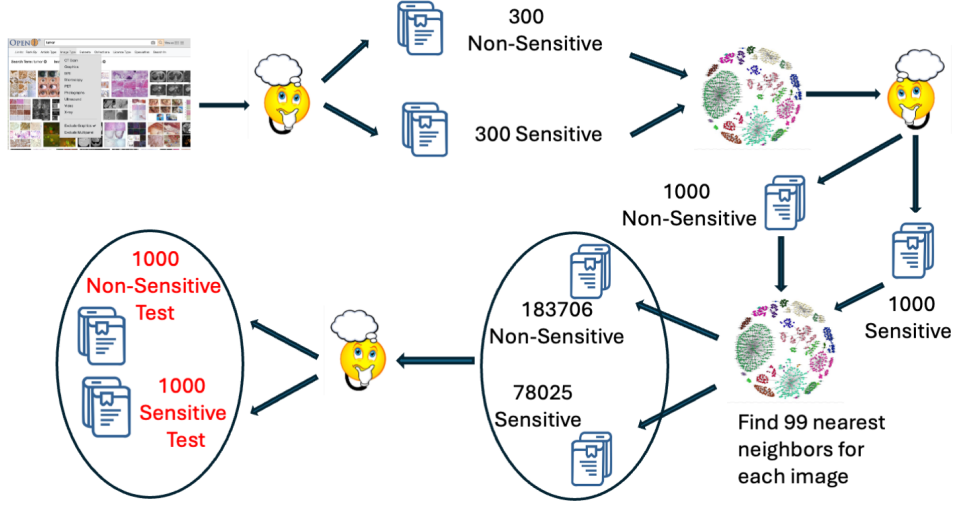


Figure 2: Illustration of MedGore dataset generation starting from 300 Open-i images to KNN based search and expansion a total of 183,706 Non-Sensitive and 78,025 Sensitive images

the probability  $\pi$  of a correct classification based on the number of correct and incorrect observations. A non-informative prior was used, with parameters  $\alpha = a+1$ ,  $\beta = b+1$  (where,  $a$  = number of NNs confirmed sensitive by human review and  $b$  = number NNs shown not-sensitive by human review) corresponding to a uniform prior distribution for  $\pi$  in the absence of prior measurements. The model has a probability density function,  $\text{PDF}(p, \alpha, \beta)$ , over possible values,  $p$ , for the actual value,  $\pi$ ,

$$\text{PDF}(p, \alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}$$

where  $B$  is the Beta function.

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

The PDF has mean  $\mu = \frac{\alpha}{\alpha+\beta}$  and variance  $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ , which is all we need.

Our estimate for  $\pi$  is,

$$\pi = \frac{\alpha}{\alpha+\beta} \pm \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$$

We evaluated three cases: combined non-sensitive and sensitive images, non-sensitive images only, and sensitive images only. For each case, we tested several sets of equal-width nearest neighbor (NN) bands. Across all configurations—irrespective of NN rank or band width—the probability of correct classification remained approximately constant. Specifically, the combined case achieved  $\sim 84\%$  accuracy, the sensitive-only case  $\sim 76\%$ , and the non-sensitive only case  $\sim 92\%$ . These findings indicate that the likelihood of an image being classified

as sensitive or non-sensitive is roughly constant out to the distance of the 100th nearest neighbor of the seed images.

To further validate this observation, we adjusted the band widths and conducted additional reviews until the uncertainty was sufficiently reduced, ensuring that the probabilities associated with each band were distinguishable. Surprisingly, the classification probability remained stable across the first 99 nearest neighbors, suggesting that the boundary between blurred and sensitive images is not reached within this range. Expanding the index to include more nearest neighbors is necessary to investigate the transition region more comprehensively.

To compare and analyze the effectiveness of the automatically generated dataset, we consider two collections of the same size as the MedGore training set from AutoMedGore for training the models. In the first set AutoMedGore-First3, we sampled images of size 2000 from the closest three nearest neighbors; in another set called AutoMedGore-Last3, we sampled images of size 2000 from the farthest three nearest neighbors. We split the set as a training and a validation set, similar to the MedGore.

### 2.3. Inter-Annotation Agreement

Before reconciling the disagreements, we analyzed inter-annotation agreement using Gwet's AC1 coefficient (Gwet, 2008) and Cohen's Kappa (Wongpakaran et al., 2013). Gwet's AC1 was shown to provide a more stable inter-rater reliability coefficient and be less affected by prevalence and marginal probability than Cohen's Kappa (Wongpakaran et al., 2013). We obtained Gwet's AC1

pairwise values ranging from 0.92 to 0.94, Cohen’s Kappa pairwise values ranging from 0.83 to 0.88 and F1 scores between 0.9 and 0.93, indicating substantial to near-perfect inter-annotator agreement and demonstrating that our annotation process is both reliable and consistent across different image categories.

## 2.4. Data Analysis

We analyzed the annotated images, which are labeled as Non-Sensitive or Sensitive based on the visual information they contained. In particular, we wanted to estimate the proportion of blood-like and skin-like regions in those images that are annotated to be blurred for public viewing. Following the literature on color detection (Chaves-González et al., 2007, 2010; Kakumanu et al., 2007), we aim to find the red color regions in the HSV color space and skin regions in the YCbCr color space. In the HSV color space, we seek the hue value corresponding to the red color, which shows the blood-like regions. We use the thresholding approach used in (Shaik et al., 2015) to detect the skin-like regions in the image. We hypothesized that if an image contains a high hue value and has been detected as a skin-like region, it is likely to be annotated as Blur. Towards this, we computed the proportions of blood-like and skin-like regions in a given image by aggregating the detected red hue scores and skin-like regions using thresholds and averaging over all values in the respective color spaces. For a given dataset, we normalized these scores using z-scores. Therefore, we assign a single estimated blur (EB) score for each image as the maximum of the blood-like and skin-like region scores. We have provided the histogram (Fig. 3) for the training and test set of MedGore, AutoMedGore-First3, and AutoMedGore-Last3. The histograms show that the average EB score (0) is mostly assigned to the Non-Sensitive images, and above average ( $>0$ ) is assigned to the Blur images. This trend is consistent across the different datasets, as shown in Fig. 3a, 3b, 3c and 3d. It also signifies that the automatically created dataset AutoMedGore-First3, and AutoMedGore-Last3 follow a similar data distribution to that of the manually assessed MedGore.

## 3. Benchmarking MedGore

### 3.1. Problem Statement

The task of sensitive medical image (SMI) classification intends to categorize a given medical image  $\mathcal{I}$  (and associated text  $\mathcal{T}$ ) into one of the classes<sup>5</sup>

{‘sensitive’, ‘non-sensitive’}. Formally, we want to learn a parametrized function  $f$  which take  $\mathcal{I}$  and  $\mathcal{T}$  (if available) and predict the class  $y = f(\mathcal{I}, \mathcal{T} : \theta)$ .

### 3.2. Approaches

We benchmark the MedGore across three settings, each reflecting a different modality configuration.

**Monomodal (Text):** We performed the experiments with the pre-trained language models such as BERT (Devlin et al., 2019) variants (BERT<sub>Base</sub>, BERT<sub>Large</sub>) and RoBERTa (Liu et al., 2019) variants (RoBERTa<sub>Base</sub>, RoBERTa<sub>Large</sub>), which are pre-trained using self-supervised objective functions on a large-scale dataset. We also utilized the recent LLMs, such as Llama3 (Grattafiori et al., 2024) models (Llama-3-8B-Instruct, Llama-3.3-70B-Instruct) and Mistral (Jiang et al., 2023) (Mistral-7B-Instruct-v0.3), which are fine-tuned on an instruction following dataset. Additionally, we included the latest Qwen (Yang et al., 2025) family models (Qwen3-8B, Qwen3-14B), which integrate the thinking mode for complex tasks requiring multi-step reasoning with the non-thinking mode for rapid, context-driven responses into a single unified framework. We fine-tune the PLMs on the training set, and for LLMs, we experiment in the zero-shot setting with the prompt shown in Fig. 4.

**Monomodal (Vision):** We utilized a variety of vision models to train them for the task of SMI classification. Specifically, we used the ResNet-50 (He et al., 2016) model (microsoft/resnet-50), which is built upon a convolutional neural network (CNN) with residual connections, and the ViT (Dosovitskiy et al., 2020) model (google/vit-base-patch16-224), a pure transformer architecture that processes images as sequences of patches, equivalent to tokens in textual data. We also employed the Swin Transformer (Liu et al., 2021) model (microsoft/swin-base-patch4-window7-224), a hierarchical Transformer that computes self-attention within shifted windows. The shifted window design makes self-attention more efficient by focusing on local windows, while still allowing connections between them. Additionally, we included recent DINOv2 (Oquab et al., 2024) models (dinov2-small, dinov2-base, dinov2-large) that are designed to stabilize and accelerate discriminative self-supervised learning as models and datasets scale.

**Multimodal (Text+Vision):** The monomodal experiments are extended to multimodal, where we experiment with a variety of models that are equipped to process textual and visual modalities to learn the complexity in information. In particular, we utilized the CLIP (Radford et al., 2021) model (clip-vit-base-patch32), which is capable of learning a shared embedding space between image and text,

<sup>5</sup>For the benchmarking purpose and reporting all the results and analysis, we will refer to the *non-sensitive*

images to be labeled as **Normal** and *sensitive* images as **Blur**.

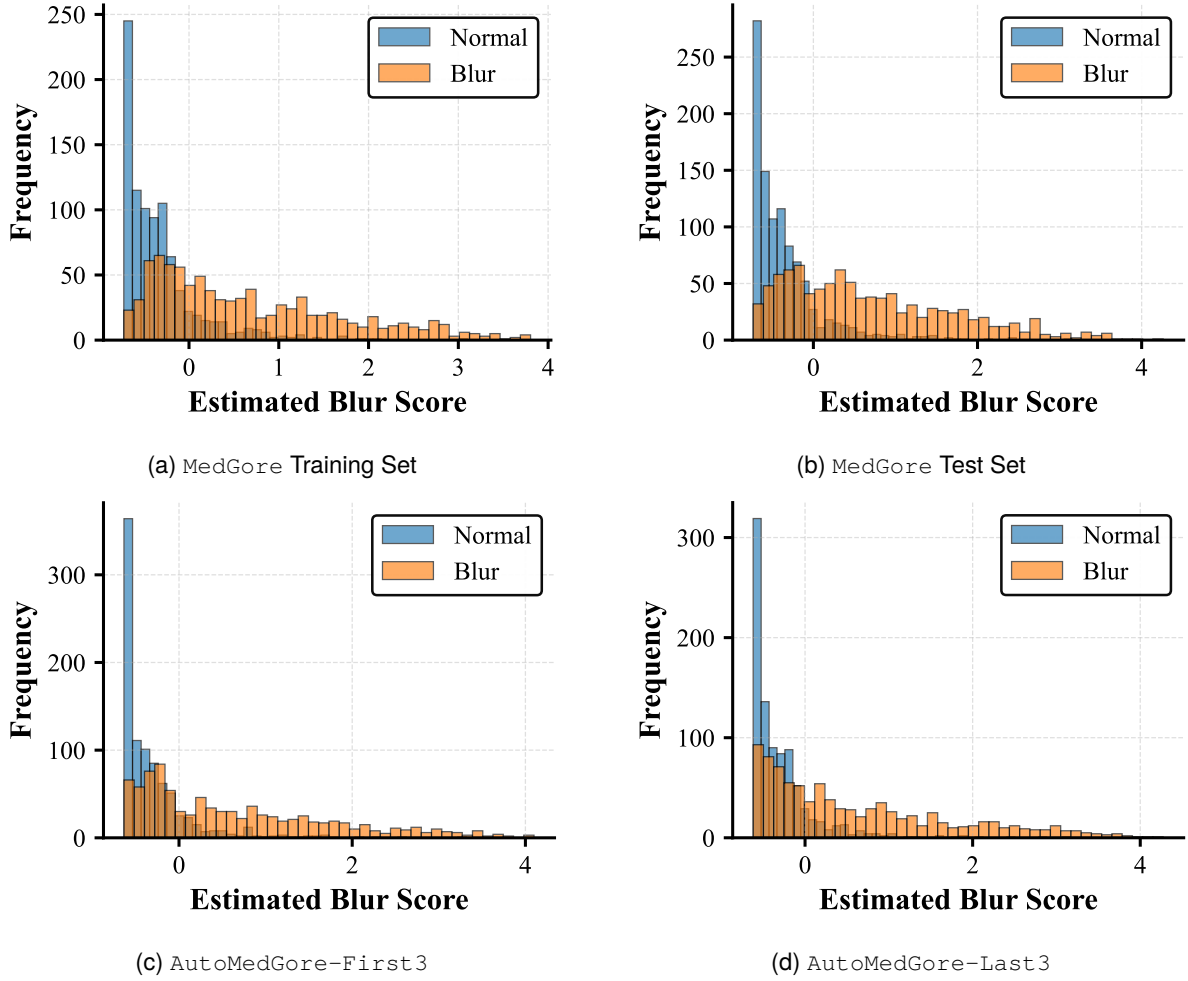


Figure 3: The estimate blur score vs. frequency of the images across the created different variants of the MedGore dataset.

#### Prompt : LLM Prompt

A medical image should be blurred if it depicts blood, tissue, organs, severe trauma, or decomposition in a manner that may be disturbing or inappropriate for general public viewing outside medical or clinical contexts. You are a classifier. Determine whether the image needs to be blurred based on the textual description of the image provided below. Respond strictly with one word: "BLUR" if the image should be blurred based on its textual description; otherwise, respond with "NORMAL".

{Textual Description}

Figure 4: Zero-shot prompt for monomodal (text) setup using large language models.

and the BLIP2 (Li et al., 2023) (blip2-flan-t5-xl) model, which leverages both frozen pre-trained image models and language models for the visual reasoning task. We also utilized the instruction-

tuned large language-vision (LLVM) models such as Qwen2-VL (Wang et al., 2024) (Qwen2-VL-7B-Instruct) model and Llava (Liu et al., 2023) model (llava-1.5-7b-hf), which are fine-tuned on large-scale instruction-following datasets to offer competitive zero-shot performance. A recent multimodal large language model (MLLM), InternVL (Zhu et al., 2025) model (InternVL3-1B-hf), is used in our experiments, which is built upon multimodal Pre-Training and mixed preference optimization and has shown superior performance across a variety of languages and tasks. All the experiments of multimodal setting are done in a zero-shot setup with the prompt shown in Fig. 5, except CLIP, where we consider the text as the candidate class labels (Normal and Blur).

### 3.3. Experimental Setups

#### 3.3.1. Datasets

We perform the experiments on human annotated image collections, MedGore and automatically

### Prompt : LLVM and MLLM Prompt

A medical image should be blurred if it depicts blood, tissue, organs, severe trauma, or decomposition in a manner that may be disturbing or inappropriate for general public viewing outside medical or clinical contexts. You are a classifier. Determine whether the image needs to be blurred based on the given textual description and the image. Respond strictly with one word: "BLUR" if the image should be blurred based on its textual description; otherwise, respond with "NORMAL".  
{Textual Description}  
{Image}

Figure 5: Zero-shot prompt for multimodal setting using large language-vision models.

generated collection, AutoMedGore-First3 and AutoMedGore-Last3. We split the set as a training and a validation set of size 1800 and 200, respectively. We train the models on the training set of MedGore. In all experiments, the validation dataset was used to select the best-performing model checkpoint, and the model was evaluated on the test set MedGore.

### 3.3.2. Modalities

**Monomodal (Text):** We consider two different sources of textual data from MedGore to perform the experiments. The first source is a **caption** of the image provided in the PubMed articles, and the second source is the **mention** sentence where the corresponding figure was referred to in the article. We trained two variants of each PLM on the training set and evaluated the model on the respective test set. Since we performed all the experiments for LLMs in a zero-shot setup, we evaluated the model on the test set, considering both **caption** and mention **textual** descriptions.

**Monomodal (Vision):** We fine-tune each vision model on the training set, considering only the image modality, and evaluate on the test dataset.

**Multimodal (Text+Vision):** Similar to the Monomodal (Text), we utilized two sets of textual descriptions to perform the experiments: caption and mention. All the experiments are done in a zero-shot setup on the test set.

### 3.4. Implementation Details

We set the maximum length of the textual description to 512 tokens for all the experiments. All the pre-trained models were used from the Hugging Face (Wolf et al., 2020) library. For the PLM's fine-tuning, we set the learning rate= $2e-5$ , epoch=10, and batch size =16. We performed the zero-shot

experiments on LLMs and VLLMs with the following configuration: maximum new tokens=4, and 4-bit quantization. For the vision models, we fine-tune them with the learning rate= $2e-5$ , epoch=10, and batch size =16.

**Evaluation Scheme** We evaluated the performance of the models in terms of macro precision, recall, and F1-score. We also computed the area under the curve score for those models that produce the probability scores for the predicted class.

## 4. Results, Analysis and Discussion

### 4.1. Results

We have reported the results with Monomodal (Text), Monomodal (Vision), and Multimodal (Text+Vision) in Tables 1, 3. For the Monomodal (Text) setting, when training with the captions of the images, the RoBERTa<sub>Large</sub> achieved the maximum F1-score of 92 on the ground-truth test set of the MedGore dataset. With the same model, we reported the F1-score of 87.35 when training with the mentions of the images. The absolute performance drop of 4.65 when switching from captions to mentions shows that captions are the most accurate representation of the images. The other models (variants of BERT and RoBERTa<sub>Base</sub>) also obtained the competitive F1-score. With the LLMs under the setting of Monomodal (Text), we obtained significantly lower performance except for the Mistral-7B-Instruct-v0.3, which reported an F1-score of 75.87 with Captions of the images and 65.65 with mentions of the images.

When considering the visual modality for the SMI classification task, we obtained better results compared to the textual modality. Under this setting, the model swin-base-patch4-window7-224 achieved the highest F1-score of 98.19, and other visual models also reported competitive performance, ranging from 94.01 (resnet-50) to 97.28 (dinov2-large). The results show that the fine-tuned visual models offer better discriminative information for the SMI classification task compared to the textual modality.

For the Multimodal (Text+Vision) setting, we conducted experiments using a zero-shot setup with five different models and reported the results in Table 1. For the clip-vit-base-patch32 model, we use the candidate labels (Blur and Normal) as text and compare the similarity with the image. The label with the highest similarity was assigned to the image. The clip-vit-base-patch32 model achieved an F1-score of 33.46 and the lowest precision score 25.15. We utilized the corresponding textual description (caption or mention) with the prompt shown in Fig. 5 for the remaining multimodal models. The model llava-1.5-7b-hf achieved



| Model Name                      | Caption   |        |          |       | Mention   |        |          |       |
|---------------------------------|-----------|--------|----------|-------|-----------|--------|----------|-------|
|                                 | Precision | Recall | F1-Score | AUC   | Precision | Recall | F1-Score | AUC   |
| <b>Monomodal (Text)</b>         |           |        |          |       |           |        |          |       |
| BERT <sub>Base</sub>            | 91.96     | 91.77  | 91.74    | 96.11 | 87.66     | 85.98  | 85.77    | 94.02 |
| BERT <sub>Large</sub>           | 92.01     | 91.92  | 91.9     | 96.72 | 88.05     | 86.23  | 86.01    | 92.77 |
| RoBERTa <sub>Base</sub>         | 91.58     | 91.28  | 91.23    | 96.27 | 87.45     | 85.78  | 85.56    | 93.86 |
| RoBERTa <sub>Large</sub>        | 92        | 92.01  | 92       | 96.82 | 88.42     | 87.47  | 87.35    | 94.56 |
| Mistral-7B-Instruct-v0.3        | 79.22     | 76.5   | 75.87    | NA    | 67.17     | 66.12  | 65.65    | NA    |
| Llama-3.3-70B-Instruct          | 50.15     | 50     | 33.65    | NA    | 73.8      | 53.19  | 40.4     | NA    |
| Llama-3-8B-Instruct             | 75.16     | 50.05  | 33.58    | NA    | 75.16     | 50.05  | 33.58    | NA    |
| Qwen3-14B                       | 37.34     | 47.22  | 34.54    | NA    | 47.3      | 49.71  | 35.36    | NA    |
| Qwen3-8B                        | 35.86     | 49.35  | 33.61    | NA    | 46.51     | 49.71  | 34.88    | NA    |
| <b>Multimodal (Text+Vision)</b> |           |        |          |       |           |        |          |       |
| clip-vit-base-patch32           | 25.15     | 50.00  | 33.46    | NA    | 25.15     | 50.00  | 33.46    | NA    |
| blip2-flan-t5-xl                | 25.52     | 31.98  | 22.42    | NA    | 17.65     | 30.66  | 22.41    | NA    |
| InternVL3-1B-hf                 | 56.70     | 50.77  | 36.92    | NA    | 58.48     | 51.43  | 38.83    | NA    |
| llava-1.5-7b-hf                 | 40.24     | 40.15  | 40.14    | NA    | 41.60     | 41.38  | 41.21    | NA    |
| Qwen2-VL-7B-Instruct            | 48.15     | 34.21  | 24.37    | NA    | 50.35     | 34.21  | 24.24    | NA    |

Table 1: Performance comparison of the monomodal (text) and multimodal (text+vision) approaches on the test set of the MedGore dataset. The ‘NA’ refers to models that generate an answer rather than predict with a confidence score.

| Model Name                                      | Caption   |        |          |       | Mention   |        |          |       |
|-------------------------------------------------|-----------|--------|----------|-------|-----------|--------|----------|-------|
|                                                 | Precision | Recall | F1-Score | AUC   | Precision | Recall | F1-Score | AUC   |
| <b>Monomodal (Text) with AutoMedGore-First3</b> |           |        |          |       |           |        |          |       |
| BERT <sub>Base</sub>                            | 84.83     | 82.87  | 82.56    | 93.8  | 86.34     | 83.29  | 82.85    | 93.35 |
| BERT <sub>Large</sub>                           | 88.78     | 86.34  | 86.06    | 96.23 | 88.03     | 86.83  | 86.67    | 94.1  |
| RoBERTa <sub>Base</sub>                         | 86.73     | 82.95  | 82.41    | 95.07 | 86.4      | 82.74  | 82.22    | 92.36 |
| RoBERTa <sub>Large</sub>                        | 88.12     | 87.22  | 87.1     | 95.78 | 87.11     | 84.69  | 84.36    | 93.09 |
| <b>Monomodal (Text) AutoMedGore-Last3</b>       |           |        |          |       |           |        |          |       |
| BERT <sub>Base</sub>                            | 86.34     | 83.29  | 82.85    | 93.35 | 86.51     | 83.84  | 83.46    | 93.2  |
| BERT <sub>Large</sub>                           | 88.03     | 86.83  | 86.67    | 94.1  | 86.57     | 84.53  | 84.25    | 93.22 |
| RoBERTa <sub>Base</sub>                         | 86.4      | 82.74  | 82.22    | 92.36 | 85.87     | 81.84  | 81.23    | 92.42 |
| RoBERTa <sub>Large</sub>                        | 87.11     | 84.69  | 84.36    | 93.09 | 87.03     | 85.28  | 85.04    | 93.73 |

Table 2: Performance comparison of the monomodal (text) approaches by training the model on AutoMedGore-First3 and AutoMedGore-Last3 on the test set of the MedGore dataset.

| Models                       | Precision | Recall | F1-Score | AUC   |
|------------------------------|-----------|--------|----------|-------|
| <b>MedGore</b>               |           |        |          |       |
| vit-base-patch16-224         | 96.87     | 96.75  | 96.73    | 96.75 |
| resnet-50                    | 94.21     | 94.03  | 94.01    | 94.03 |
| swin-base-patch4-window7-224 | 98.2      | 98.19  | 98.19    | 98.19 |
| dinov2-base                  | 96.44     | 96.43  | 96.43    | 96.43 |
| dinov2-large                 | 97.34     | 97.29  | 97.28    | 97.29 |
| dinov2-small                 | 96.15     | 96     | 95.97    | 96    |
| <b>AutoMedGore-First3</b>    |           |        |          |       |
| vit-base-patch16-224         | 94.52     | 94.05  | 94       | 94.05 |
| resnet-50                    | 89.8      | 88.09  | 87.91    | 88.09 |
| swin-base-patch4-window7-224 | 94.96     | 94.55  | 94.51    | 94.55 |
| dinov2-base                  | 91.21     | 89.85  | 89.71    | 89.85 |
| dinov2-large                 | 92.46     | 91.35  | 91.24    | 91.35 |
| dinov2-small                 | 91.02     | 90.58  | 90.52    | 90.58 |
| <b>AutoMedGore-Last3</b>     |           |        |          |       |
| vit-base-patch16-224         | 94.32     | 93.89  | 93.85    | 93.89 |
| resnet-50                    | 92.12     | 91.05  | 90.94    | 91.05 |
| swin-base-patch4-window7-224 | 93.81     | 93.2   | 93.14    | 93.2  |
| dinov2-base                  | 91.96     | 90.94  | 90.84    | 90.94 |
| dinov2-large                 | 92.01     | 90.65  | 90.52    | 90.65 |
| dinov2-small                 | 92.48     | 91.74  | 91.67    | 91.74 |

Table 3: Performance comparison of the monomodal (vision) approaches by training the model on MedGore, AutoMedGore-First3 and AutoMedGore-Last3 on the test set of the MedGore dataset.

the highest F1-score of 40.14 amongst all multimodal models with the balanced precision and recall scores. The model InternVL3-1B-hf obtained the highest precision score of 56.70 and recall score of 50.77. The model predicted most of the 1929 images into the Blur category out of a total of 1988 images in the test set, which yielded a high F1-score (66.9) of class Blur but a very low F1-score (7.1) for Normal images, leading to the lower macro-F1 score of 36.92.

## 4.2. Analysis

To analyze the effectiveness of the automatically created dataset, we conducted additional experiments, selecting samples using two different strategies (AutoMedGore-First3 and AutoMedGore-Last3) to create a training set for comparison with the human-annotated training set MedGore. The experimental results depicted in Tables 1, 3 show that when the model was trained on the training set of MedGore, it performed bet-



ter in monomodal (text) and monomodal (vision) settings. Following this, we aim to analyze the effect of model performances when they are trained on `AutoMedGore-First3` and `AutoMedGore-Last3` datasets. For the monomodal (text) setting, we reported the results in Table 2 by training the models with `AutoMedGore-First3` and `AutoMedGore-Last3` datasets and evaluating on the same ground-truth dataset, which is part of `MedGore` collection. We obtained the best results with the `RoBERTaLarge` model, achieving an F1-score of 87.1 and 84.36 with `AutoMedGore-First3` considering the caption and mention, respectively, as the textual description of the images. Comparing these results with the `MedGore` training set, we can say that though there is a performance decrement, it has also to be noted that the `AutoMedGore-First3` is our automatically created dataset, which is achieving stronger results compared to the resource and time-consuming construction of the human-annotated `MedGore` dataset. We observe similar trends when the `AutoMedGore-Last3` was considered for training the models. We extend the same experiments with a monomodal (vision) setting and reported and compared the results in Table 3. We observe that the best-performing `swin-base-patch4-window7-224` model achieved an F1-score of 94.51 and 93.14 for `AutoMedGore-First3` and `AutoMedGore-Last3`, respectively. These results allow us to make two claims: (1) our automatically created dataset is producing comparable results for the task of SMI classification, (2) the quality of the automatically generated datasets remains similar as we descend into nearest neighbor ranking while choosing the samples.

### 4.3. Discussion

The analysis of NN homogeneity estimation using the Beta-Binomial Bayesian Model demonstrates that clusters formed around the seed images are consistent and homogeneous. The probability of correct classification remains fairly constant at approximately 84% for distances out to the 100th NN, indicating that the retrieved images are visually very similar to their corresponding seeds. A higher level of homogeneity is observed for non-sensitive images ( $\sim 92\%$ ) compared to sensitive images ( $\sim 76\%$ ), likely due to the greater number and visual uniformity of the non-sensitive seeds locally. Overall, these results suggest that clusters within the first 99 NNs are well separated and maintain strong class consistency.

The KNN expansion resulted in 183,000 non-sensitive images and 78,000 sensitive images. This imbalance arises because 77.45% of non-sensitive images are unique, whereas only 31.85% of sensitive images are unique. This indicates that

sensitive image clusters exhibit greater overlap, while non-sensitive image clusters are more widely distributed across the feature space. The higher redundancy among sensitive images can be attributed to their visual similarity in image features, such as the predominance of red tones in surgical photographs. Additionally, a small proportion of non-photographic images, such as MRI or X-ray scans, are included in the non-sensitive category. Also some photographs were also filtered during processing, further contributing to this distributional difference.

The experiments on a variety of monomodal and multimodal approaches reveal that the task of SMI classification is well handled by a hierarchical vision transformer (Dosovitskiy et al., 2020), which processes images in patches. However, `ResNet50` achieved the highest performance, even though it was not one of the computationally inexpensive models evaluated in our study. We also find that the monomodal (text) model with textual descriptions (captions or mentions) as input for images is not better than the monomodal (vision) model. It can also be attributed to our visual analysis (see Section 2.4), which shows that images with blood-like and skin-like regions tend to produce higher estimated blur scores. We observed around 7% performance decrement between the best-performing monomodal (text) and monomodal (vision) models. In our experiments, we found that neither LLMs nor VLLMs under a zero-shot setting could perform better on the SMI classification task. Given the better performance of the lightweight vision model, we did not extend our study to fine-tune those LLMs and VLLMs, which require substantial computational resources, as this would not be a practical solution for model deployment. We also observe that the high performance of the SMI classification could be due to a binary classification setup, which makes the model learn discriminative features between normal and gory images. A more challenging problem could be fine-grained gory image classification, where the model needs to identify a specific type of gory image, if any is present. We will consider this extension in future work.

## 5. Conclusion

In this work, we introduce a large dataset of medical photographs, their captions and corresponding discussions in the full text of the medical articles from which the images were extracted. The dataset was generated to enable development of approaches to automated detection of sensitive medical images. We introduce a KNN-based approach to rapid labeling of the images and an approach to NN homogeneity estimation using the Beta-Binomial Bayesian Model. Finally, we establish mono- and multi-modal benchmarks for image

labeling using image and text features.

## 6. Ethical Considerations and Limitations

The images in the MedGore collection are sourced from the Open Access subset of PubMedCentral. The dataset may be freely downloaded, reused, redistributed, and publicly shared under the applicable open-access license. To the best of our knowledge, the journals that publish the images require patients' consent and IRB review. These images are publicly available at <https://openi.nlm.nih.gov>. While the dataset we presented is large, covers multiple aspects of sensitive image content, and enables highly accurate image classification, our analysis shows that some unknown sensitive image types may be present in Open-i and other medical image collections. We believe that the presented approach to the identification of such images will enable filling these gaps in future work.

## 7. Data Availability

MedGore Dataset is available for download at [https://bionlp.nlm.nih.gov/MedGore\\_Dataset.zip](https://bionlp.nlm.nih.gov/MedGore_Dataset.zip).

## 8. 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.

## 9. Bibliographical References

Fatih Cagatay Akyon and Alptekin Temizel. 2023. [State-of-the-art in nudity classification: A comparative analysis](#). In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.

Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. 2020. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283.

Matthias Carstens, Franziska M Rinner, Sebastian Bodenstedt, Alexander C Jenke, Jürgen Weitz, Marius Distler, Stefanie Speidel, and Fiona R Kolbinger. 2023. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10(1):1–8.

Jose M Chaves-González, Miguel A Vega-Rodríguez, Juan A Gómez-Pulido, and Juan M Sánchez-Pérez. 2007. Colour spaces study for skin colour detection in face recognition systems. In *International Conference on Security and Cryptography*, volume 2, pages 171–174. SCITEPRESS.

Jose M Chaves-González, Miguel A Vega-Rodríguez, Juan A Gómez-Pulido, and Juan M Sánchez-Pérez. 2010. Detecting skin in face recognition systems: A colour spaces study. *Digital signal processing*, 20(3):806–823.

Dina Demner-Fushman, Sameer Antani, Matthew Simpson, and George R Thoma. 2012. Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, 6(2):168–177.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

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*.

Deepak Gupta, Russell Loane, Soumya Gayen, and Dina Demner-Fushman. 2023. Medical image retrieval via nearest neighbor search on pre-trained image features. *Knowledge-based systems*, 278:110907.

Kilem L. Gwet. 2008. *Computing inter-rater reliability and its variance in the presence of high agreement*. Advanced Analytics LLC.

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7b*.
- Praveen Kakumanu, Sokratis Makrogiannis, and Nikolaos Bourbakis. 2007. A survey of skin-color modeling and detection methods. *Pattern recognition*, 40(3):1106–1122.
- Sasan Karamizadeh, Saman Shojae Chaeikar, and Alireza Jolfaei. 2023. Adult content image recognition by boltzmann machine limited and deep learning. *Evolutionary Intelligence*, 16(4):1185–1194.
- William Larocque. 2021. *Gore classification and censoring in images*. Ph.D. thesis, Universit   d’Ottawa/University of Ottawa.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- 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*.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.
- Maxime Oquab, Timoth  e Darcet, Th  o Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. *DINOv2: Learning robust visual features without supervision*. *Transactions on Machine Learning Research*. Featured Certification.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. *Learning transferable visual models from natural language supervision*.
- Manuel Sebasti  n R  os, Mar  a Alejandra Molina-Rodr  guez, Daniella Londo  o, Camilo Andr  s Guill  n, Sebasti  n Sierra, Felipe Zapata, and Luis Felipe Giraldo. 2023. Cholec80-cvs: An open dataset with an evaluation of strasberg’s critical view of safety for ai. *Scientific Data*, 10(1):194.
- Khamar Basha Shaik, P Ganesan, V Kalist, BS Sathish, and J Merlin Mary Jenitha. 2015. Comparative study of skin color detection and segmentation in hsv and ycbcr color space. *Procedia Computer Science*, 57:41–48.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. *Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution*. *CoRR*, abs/2409.12191.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 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.
- Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen’s kappa and gwet’s ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC medical research methodology*, 13(1):61.
- 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*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. InternV3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

# Diagnostic Reasoning with Large Language Models for a Rare Disease: Case Study of Primary Ciliary Dyskinesia

Swati Rajwal<sup>1</sup>, Mary Ellen Fain<sup>2,3</sup>, Lokesh Guglani<sup>2,3</sup>, Abeed Sarker<sup>1</sup>

<sup>1</sup>Dept. of Biomedical Informatics, Emory University, <sup>2</sup>Dept. of Pediatrics, Emory University

<sup>3</sup> Children's Healthcare of Atlanta

{swati.rajwal, mary.ellen.fain, lokesh.guglani, abeed.sarker}@emory.edu

## Abstract

Primary ciliary dyskinesia (PCD) is a rare pediatric lung disease that is frequently underdiagnosed due to nonspecific early symptoms and limited clinical exposure. We investigate whether large language models (LLMs) can support early diagnostic reasoning using real-world pediatric pulmonology notes written before the final diagnosis. We curated 58 de-identified first-visit notes (28 confirmed PCD, 30 controls) and evaluated five open-source LLMs using a standardized zero-shot prompt to produce structured outputs, including PCD evaluation recommendations, justifications, and suggested tests. Quantitative performance was assessed against expert-validated labels using sensitivity, specificity, and accuracy, and a clinician qualitatively reviewed all explanations and testing recommendations for clinical soundness. Sensitivity ranged from 0.48 to 1.00 and specificity from 0.10 to 0.48 (excluding uncertain outputs), with a best accuracy of 0.75. A majority-vote ensemble of five open-source LLMs achieved perfect sensitivity (1.00) with accuracy of 0.73. While models often identified clinically relevant signals in unstructured notes, explanations and testing recommendations were frequently only partially sound. These findings suggest LLMs may serve as cautious early screening aids for rare disease suspicion, but not as standalone diagnostic tools. This work further highlights the need for larger, multi-site evaluation on longitudinal clinical text.

**Keywords:** Large Language Models, Primary Ciliary Dyskinesia, Rare Disease, Natural Language Processing

## 1. Introduction

Primary ciliary dyskinesia (PCD) is a rare inherited condition that affects the normal movement of cilia in the airways and other organs. Children with PCD often experience chronic cough, recurrent respiratory infections, ear and sinus disease, and in some cases organ laterality differences. Despite these characteristic features, PCD is frequently diagnosed late or underdiagnosed (Salinas et al., 2025). There is no single definitive test for PCD (Behan et al., 2016; Knowles et al., 2016), and diagnosis depends on combining clinical history with several specialized investigations (Zariwala et al., 1993–2025). Many children are evaluated multiple times before PCD is seriously considered, leading to delays in care and preventable lung damage.

These diagnostic challenges occur in the context of limited clinical exposure. Most clinicians encounter very few PCD cases during training or practice, and symptoms overlap with much more common childhood respiratory conditions. As a result, PCD may not be included early in the diagnostic reasoning process, even when classic features are present, a challenge that has prompted the development of several international registries and multi-center cohorts in the field (PCD Foundation, 2026; PCD Support UK, 2026). Prior work has shown that expert review of large collections of electronic medical records (EMRs) can help identify patients with previously missed PCD diagnoses,

highlighting how early clinical clues are often documented but not acted upon (Fecho et al., 2022; Zhou et al., 2025).

Large language models (LLMs) are now widely accessible to clinicians, patients, and families. These systems can read and summarize long clinical narratives, recognize symptom patterns, and generate structured responses to medical inquiries (Wei et al., 2024; Das et al., 2025) or errors (Rajwal et al., 2024). Parents report using such systems to ask whether their child's symptoms could indicate a rare condition,<sup>1</sup> and clinicians increasingly report LLM use as decision support tools (Tausky et al., 2025). However, most evidence on LLM performance comes from general medical knowledge tests or synthetic cases, rather than real-world clinical notes written before a diagnosis is known.

To fill this gap in the literature, we investigate a clinically important question: when presented with real-world pediatric pulmonology notes, do LLMs exhibit clinically appropriate diagnostic reasoning for PCD? What systematic failure modes emerge? We use real, de-identified pediatric clinical notes from patients who were evaluated for PCD, including both confirmed cases and controls in whom PCD was ruled out. Overall, this study provides an early assessment of how LLMs may behave when applied to real pediatric diagnostic reasoning for a

<sup>1</sup>[today.com/health/mom-chatgpt-diagnosis-pain-rcna101843](https://today.com/health/mom-chatgpt-diagnosis-pain-rcna101843). Accessed on 04/05/2026



rare respiratory disease<sup>2</sup>. Our key contributions:

- We evaluate how multiple open-source LLMs reason about suspected PCD using real first-visit pediatric pulmonology notes collected before the final diagnosis.
- We conduct quantitative evaluation against expert-validated gold-standard labels using sensitivity, specificity, and accuracy.
- We perform clinician-led qualitative review of models' explanations and PCD testing recommendations to assess clinical reasoning.

## 2. Related Work

Diagnosing primary ciliary dyskinesia (PCD) remains inherently complex, with studies showing that different guideline-based diagnostic algorithms can yield contradictory diagnoses in a significant subset of patients (Nussbaumer et al., 2021). This diagnostic uncertainty has motivated growing interest in computational approaches that support rare disease recognition. Recent studies have explored the use of large language models (LLMs) in the context of rare diseases (Hodges et al., 2025), motivated by the long diagnostic delays and limited specialist availability that many patients face (Chen et al., 2024; Shyr et al., 2024; Abdullahi et al., 2024; Wu et al., 2024; Hodges et al., 2025). Several works have shown that LLMs can support rare disease diagnosis, education (Niyonkuru et al., 2025), and clinical reasoning by synthesizing information across symptoms, histories, and medical literature. Benchmarks such as RareBench (Chen et al., 2024) and qualitative evaluations on complex diagnostic cases suggest that general-purpose LLMs are often able to include rare diseases in their diagnostic considerations, even when these conditions are infrequently encountered in routine clinical practice.

Several studies have further demonstrated that zero-shot LLMs, which do not require task-specific fine-tuning, can perform comparably to more complex pipelines for rare disease tasks such as generating plausible differential diagnoses or extracting clinically relevant information from unstructured clinical text (Wu et al., 2024; Shyr et al., 2024; Ríos-Hoyo et al., 2024). These findings motivate our use of a zero-shot setting, which better reflects real-world constraints and avoids reliance on curated training datasets that are difficult to obtain for rare conditions.

In contrast to the growing literature on rare diseases broadly, there has been relatively little work

focused specifically on PCD. Existing studies related to PCD have primarily examined how traditional machine learning or information retrieval approaches (Rother et al., 2015; Burns et al., 2025) can be used to identify patients who may have been missed during routine clinical care (Zhou et al., 2025). For example, prior work has shown that searching EMRs for characteristic symptom patterns can help uncover patients with undiagnosed PCD, highlighting the presence of early clinical signals that are often documented but not recognized at the time (Zhou et al., 2025). However, these approaches do not directly evaluate how language-based models reason over individual clinical notes in a diagnostic context. While these studies demonstrate the feasibility of identifying potential PCD cases using structured data or retrieval methods, our work shifts the focus to LLM-based reasoning using unstructured clinical notes, which more closely reflects real-world diagnostic settings.

## 3. Methods

Figure 1 shows the pipeline used in this study. Below, we explain the dataset, LLMs used, and their implementation details.

### 3.1. Data

This study uses de-identified clinical notes from pediatric patients seen at Children's Hospital of Atlanta (CHOA), collected under an approved Institutional Review Board (IRB) protocol (#STUDY00002645). The dataset includes two groups: children who were later confirmed to have PCD [ $n=28$ , mean age= 4.85 years ( $\pm 5.94$ ), sex= 12 M, 16 F] and children who were evaluated for PCD but ultimately did not have the disease [ $n=30$ , mean age= 7.58 years ( $\pm 1.87$ ), sex= 13 M, 17 F] to serve as controls. **Confirmed PCD cases** were identified via a pre-existing list maintained in the EMRs of children with PCD being cared for at CHOA, which included 44 patients. Manual chart review was performed for each patient to identify the first clinical encounter with a pediatric pulmonologist at the institution, to reflect the child's initial symptom onset and diagnostic evaluation before PCD was confirmed or ruled out. Patients were excluded if they were referred with a pre-existing diagnosis of PCD or if their first clinical encounter available in the EMRs followed a confirmatory diagnosis ( $n=11$ ), as this would remove diagnostic uncertainty.

**Control cases** were selected from a broader pre-existing list in the EMRs of patients referred to the PCD specialty clinic, which was cross-referenced with the confirmed PCD list to identify patients in whom PCD was suspected but not confirmed. Manual chart review was similarly performed. Patients

<sup>2</sup>Preliminary version of this work was accepted at the ATS 2026 International Conference (Rajwal et al., 2026).

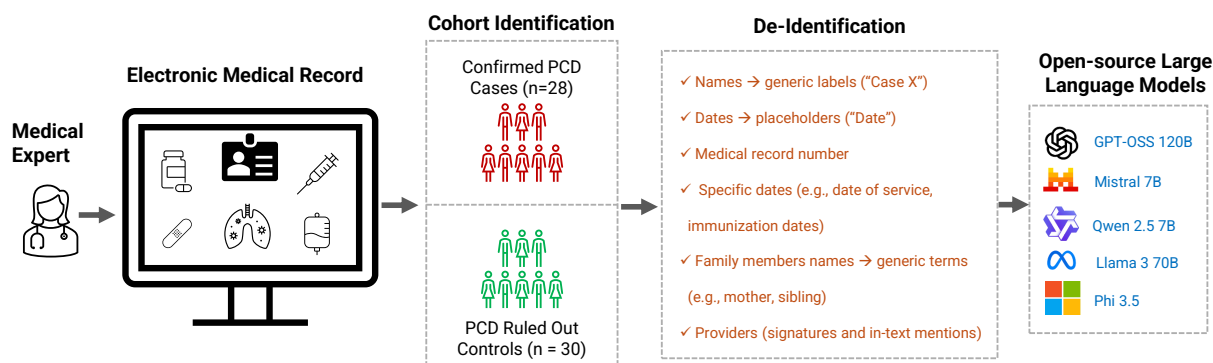


Figure 1: Study pipeline showing cohort curation and de-identification of first-visit clinical notes, zero-shot inference with five open-source LLMs. Subsequently, quantitative evaluation and clinician-led qualitative review of model outputs are reported in this study.

were included if their workup was ultimately negative for PCD, and excluded if their workup was still ongoing with uncertain diagnosis.

For the selected patients, the identifying information was removed from their encounter note such as medical record number, date of birth, names (of the patient, replaced by "Case" and a case number, as well as family member or provider names), and specific dates. Each selected note was used in its entirety as input to the LLMs and for expert annotation, without additional filtering of clinical content beyond standard de-identification.

### 3.2. Large Language Models Used

We evaluated five open-source LLMs: GPT-OSS-120B (Agarwal et al., 2025), Llama-3-70B (AI@Meta, 2024), Qwen-2.5-7B (Yang et al., 2024), Mistral-7B (Mistral AI, 2024), and Phi-3.5-mini (Abdin et al., 2024). We chose these models to reflect diversity in open-source LLM design and scale. Their differing behaviors are consistent with known trade-offs, such as higher sensitivity in smaller models and more conservative reasoning in larger ones. Each model was given the same zero-shot prompt instructions to assess a single clinical note per patient. Open-source models were used to enable local inference on secure, regulatory-compliant university servers without transmitting clinical data to external systems.

We utilized a zero-shot prompting approach to produce structured JSON output containing: (1) a recommendation for further PCD evaluation (Yes/No/Uncertain), (2) justification for that recommendation, (3) recommended diagnostic tests, and (4) justification for testing decisions. Below, we present the prompt structure:

#### Instruction Header

You are a clinician. Follow the schema exactly. Be concise and deterministic. Rules:  
- Output JSON with exactly these keys:

```
"pcd_recommendation", "justification_pcd",
"testing_recommended",
"justification_tests".
- "pcd_recommendation" must be "Yes", "No", or
"Uncertain".
- Each justification in around 90 words.
- No extra text. After JSON, write END_OF_OUTPUT.
...
```

#### Key Meanings

```
...
- "pcd_recommendation": Clinical impression -
does this patient need further PCD evaluation?
(Yes/No/Uncertain)
- "justification_pcd": Brief reasoning citing find-
ings from the note.
- "testing_recommended": Diagnostic tests you
would order next. Use "None" if not needed.
- "justification_tests": Why those tests (or why
none).
...
```

#### Sharing Clinical Note

```
...
Below is a de-identified clinical note from a patient's
first pulmonology encounter. Read it and respond per
the schema.
[NOTE STARTS]
{text}
[NOTE ENDS]
...
```

#### Output Formatting Instructions

```
...
Reply ONLY in the following JSON structure:
{{
 "pcd_recommendation": " ",
 "justification_pcd": " ",
 "testing_recommended": " ",
 "justification_tests": " "
}}
END_OF_OUTPUT
```

In addition to individual model evaluations, we constructed an ensemble using a majority voting strategy across the five LLMs used in this study. For each clinical note, the final ensemble prediction was defined as the most frequently predicted label (Yes/No/Uncertain) across models. In cases of ties, the prediction was resolved as “Uncertain.”

### 3.3. Implementation Details

Open-source LLM weights were downloaded from the HuggingFace Model Hub and stored locally on University server. Entire code<sup>3</sup> was implemented in Python 3.10.15. Experiments were executed using SLURM with a consistent hardware configuration of one NVIDIA Ampere GPU, 3 CPU cores, and 80 GB of system memory. Across models, inference time ranged from 5 to 15 minutes per 30 cases, with an average runtime of about 10 minutes.

### 3.4. Evaluation & Manual Analysis

We evaluate models using clinically meaningful metrics, including sensitivity and specificity for recommending further PCD evaluation. We report 95% confidence intervals (CIs) for all metrics, computed using the Wilson score interval for binomial proportions (sensitivity, specificity, accuracy) and non-parametric bootstrapping with 10,000 resamples ( $F_1$ -score). We also perform manual annotation to review LLM outcomes. One PCD clinician based at CHOA manually reviewed the LLM output using an annotation guideline<sup>4</sup> (co-developed with a second PCD clinician) for the justification of the PCD recommendation, as well as any suggested diagnostic tests and their explanations. The annotator rated diagnostic explanations as clinically sound, partially sound, clinically weak, or clinically concerning, and noted any hallucinations. Suggested diagnostic tests were labeled as appropriate, incomplete, excessive, inappropriate, or potentially harmful. Explanations for test recommendations were separately rated based on whether they clearly justified the suggested tests. The annotator was blinded to model identity and based all judgments only on the provided note and model output.

## 4. Results

Figure 2 presents the distributions of model responses, illustrating clear differences in how models recommend further evaluation for PCD across confirmed and non-PCD cases. Models were more likely to recommend further evaluation for confirmed cases, though they varied in how strongly they favored a positive recommendation versus an

uncertain one. Mistral-7B and Phi-3.5-mini showed the strongest tendency toward recommending evaluation, identifying nearly all confirmed cases, while GPT-OSS-120B produced substantially fewer positive recommendations. The ensemble (majority voting) balanced these behaviors, recommending evaluation for most confirmed cases while reducing extreme responses seen in individual models. Figure 3 shows agreements between the models.

Table 1 summarizes performance excluding uncertain responses. The number of uncertain responses varied across models, ranging from 2/58 (3.4%) for GPT-OSS-120B to 17/58 (29.3%) for Llama-3-70B. Sensitivity ranged from 0.48 to 1.00, whereas specificity remained below 0.50 for all models. Phi-3.5-mini achieved the highest overall accuracy (0.75), followed closely by Qwen 2.5 7B (0.70). The ensemble (majority-vote) achieved perfect sensitivity (1.00) with an accuracy of 0.73, matching the strongest individual models in identifying confirmed cases but without improving specificity beyond that range. We additionally report a stricter setting where *Uncertain* predictions are treated as incorrect. Figure 4 shows examples of hallucinations. The expert annotator’s evaluation of LLM responses is shown in Figure 5.

## 5. Discussion

### 5.1. Diagnostic Performance

Across models, the dominant pattern is high sensitivity paired with weak specificity, consistent with a screening-oriented behavior rather than definitive decision-making. Although this is expected, given that all patients were referred for PCD evaluation, it reflects a consistent tendency of the models to favor sensitivity over specificity in clinically ambiguous settings. When uncertain predictions are excluded, sensitivity ranges from 0.48 to 1.00, while specificity remains low for all models (0.095 to 0.483). The highest accuracy observed is 0.750 (Phi-3.5-mini), and the ensemble achieves 0.727 with sensitivity of 1.000, but specificity remains limited at 0.429. These results indicate that while models rarely miss confirmed cases under this evaluation, they frequently flag non-PCD cases for further evaluation, which could increase unnecessary testing or referrals if outputs are used without clinical oversight.

Figure 2 and Table 1 together suggest that the “Uncertain” option plays a meaningful role in how models manage risk. Models vary in how often they select this option, and these differences influence the apparent trade-off between missed cases and false positives. Performance metrics computed after excluding “Uncertain” therefore reflect behavior conditional on a model choosing to make a decision,

<sup>3</sup>GitHub Code: [pcd-diagnosis-llm-study](#)

<sup>4</sup>Annotation Guideline: [Annotation-Guideline.pdf](#)

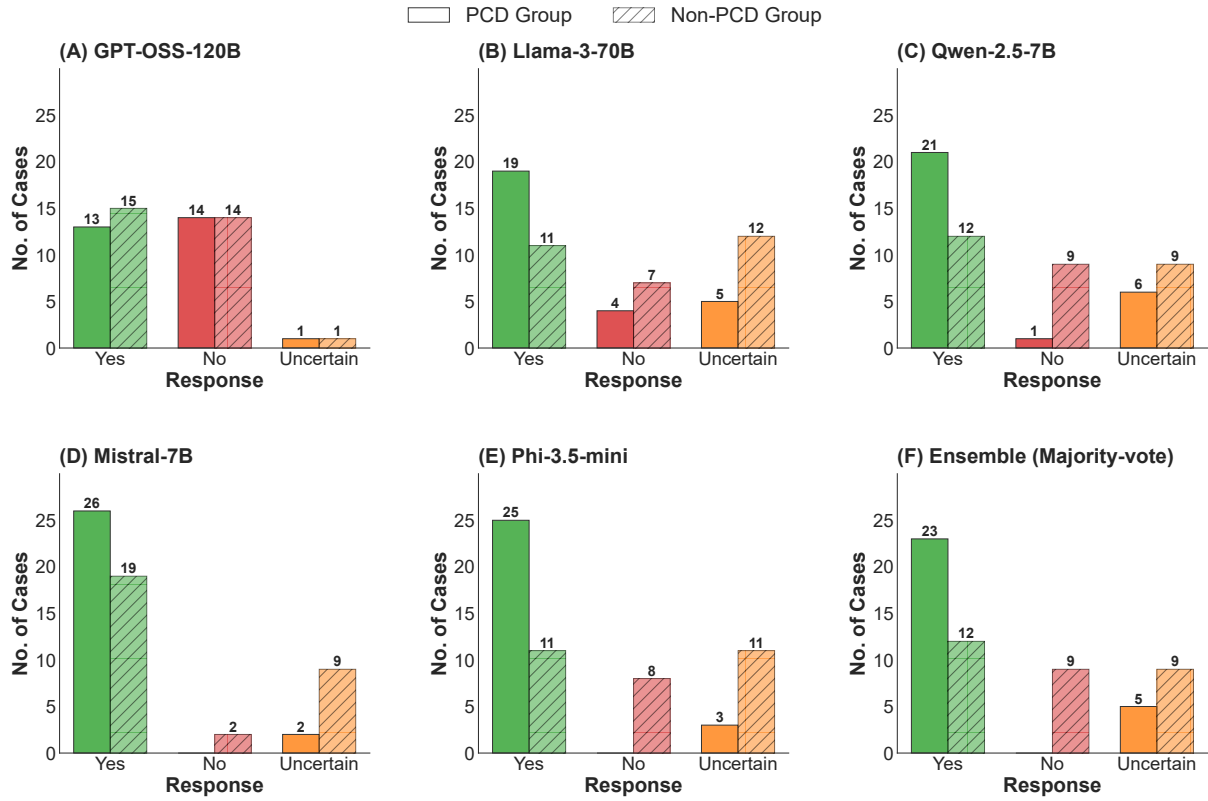


Figure 2: Distribution of model recommendations ("Yes", "No", "Uncertain") across confirmed PCD cases (n=28) and non-PCD controls (n=30) for each LLM and a majority-vote ensemble of five models. Counts are shown separately for each group.

| Model                                                        | Evaluated Cases<br>(Abstention Rate) | Sensitivity          | Specificity          | Accuracy             | F <sub>1</sub> Score |
|--------------------------------------------------------------|--------------------------------------|----------------------|----------------------|----------------------|----------------------|
| Approach 1: Committed cases only (uncertain excluded)        |                                      |                      |                      |                      |                      |
| GPT-OSS-120B                                                 | 56 (3.4%)                            | 0.481 [0.307, 0.660] | 0.483 [0.314, 0.656] | 0.482 [0.357, 0.610] | 0.473 [0.314, 0.613] |
| Llama-3-70B                                                  | 41 (29.3%)                           | 0.826 [0.629, 0.930] | 0.389 [0.203, 0.614] | 0.634 [0.481, 0.764] | 0.717 [0.600, 0.824] |
| Qwen-2.5-7B                                                  | 43 (25.9%)                           | 0.955 [0.782, 0.992] | 0.429 [0.245, 0.635] | 0.698 [0.549, 0.814] | 0.764 [0.678, 0.851] |
| Mistral-7B                                                   | 47 (19.0%)                           | 1.000 [0.871, 1.000] | 0.095 [0.027, 0.289] | 0.596 [0.453, 0.724] | 0.732 [0.667, 0.800] |
| Phi-3.5-mini                                                 | 44 (24.1%)                           | 1.000 [0.867, 1.000] | 0.421 [0.231, 0.637] | 0.750 [0.606, 0.854] | 0.820 [0.746, 0.893] |
| Ensemble (Majority-vote)                                     | 44 (24.1%)                           | 1.000 [0.857, 1.000] | 0.429 [0.245, 0.635] | 0.727 [0.582, 0.837] | 0.793 [0.719, 0.868] |
| Approach 2: Pessimistic bound (uncertain = wrong prediction) |                                      |                      |                      |                      |                      |
| GPT-OSS-120B                                                 | 58                                   | 0.464 [0.295, 0.642] | 0.467 [0.302, 0.639] | 0.466 [0.343, 0.592] | 0.456 [0.298, 0.596] |
| Llama-3-70B                                                  | 58                                   | 0.679 [0.493, 0.821] | 0.233 [0.118, 0.409] | 0.448 [0.327, 0.575] | 0.543 [0.424, 0.656] |
| Qwen-2.5-7B                                                  | 58                                   | 0.750 [0.566, 0.873] | 0.300 [0.167, 0.479] | 0.517 [0.392, 0.641] | 0.600 [0.486, 0.703] |
| Mistral-7B                                                   | 58                                   | 0.929 [0.774, 0.980] | 0.067 [0.018, 0.213] | 0.483 [0.359, 0.608] | 0.634 [0.564, 0.709] |
| Phi-3.5-mini                                                 | 58                                   | 0.893 [0.728, 0.963] | 0.267 [0.142, 0.444] | 0.569 [0.441, 0.688] | 0.667 [0.582, 0.750] |
| Ensemble (Majority-vote)                                     | 58                                   | 0.821 [0.644, 0.921] | 0.300 [0.167, 0.479] | 0.552 [0.425, 0.673] | 0.639 [0.541, 0.732] |

Table 1: Performance with 95% CI. Approach 1 excludes *Uncertain* cases (committed predictions only, i.e., yes/no). Abstention rate = % of *Uncertain* predictions; Approach 2 treats *Uncertain* as wrong prediction.

rather than end-to-end system behavior. In an unstructured clinician review of cases labeled as "uncertain," the underlying source of uncertainty was noted to be variable, with some labels seeming to reflect an incomplete discussion of PCD features in the source note, some reflecting the source note's suspicion for an alternate diagnosis, and still others directly referring to inconclusive PCD diagnostic

testing (for example, genetic variants of uncertain significance). In practice, any clinical workflows need explicit guidance on how to handle uncertain outputs, as they represent a substantial fraction of model responses. Additional structured review of these cases in the future may help identify patterns that could be leveraged to minimize uncertain output, or at minimum help interpret this response.



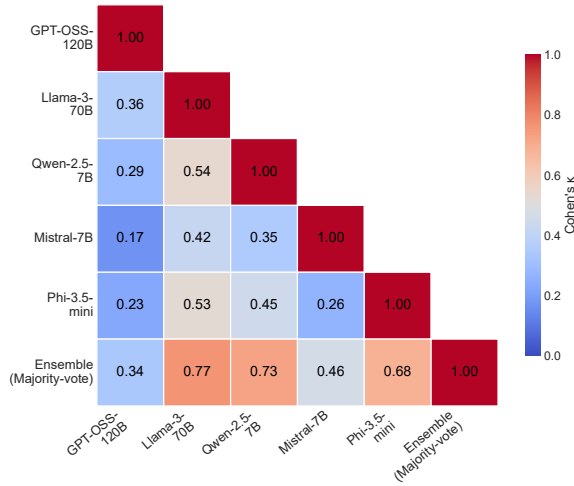


Figure 3: Inter-model agreement for PCD recommendation heatmap across all cases (N=58).

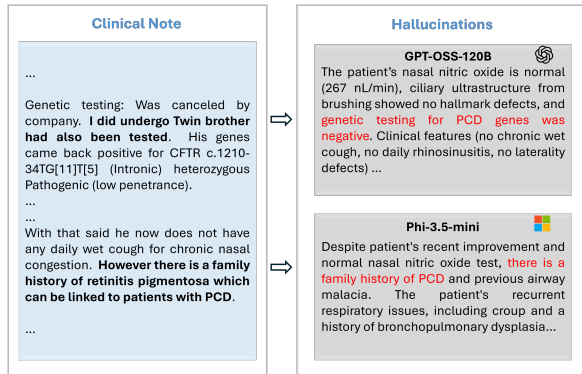


Figure 4: Hallucinations in PCD justification by two LLMs. Note that the part of clinician's note shown is for reference. The text marked in red shows misinterpretation by model.

## 5.2. Qualitative Performance

Manual annotations by domain experts and further analyses highlight that factual grounding does not imply strong clinical reasoning (Fig. 5). This pattern is consistent with prior work showing that LLMs can produce correct answers while generating inconsistent reasoning (Rajwal et al., 2025). Hallucination rates were low across models (0.0–8.6%), yet "clinically sound" diagnostic explanations accounted for only 20.7–48.3% of outputs, while "partially sound" explanations consistently represented 34.5–44.8%. "Clinically weak" explanations reached as high as 37.9% for some models. Notably, models with zero hallucinations, such as Llama-3-70B and Mistral-7B, still produced "clinically weak" or "concerning" explanations in a substantial fraction of cases, showing that avoiding unsupported statements is not sufficient for reliable clinical reasoning.

A large portion of clinically weak labels were applied due to incongruous outputs for further PCD

evaluation recommendation (Yes/No/Uncertain) and the justification of that answer. Specifically, models frequently answer "Yes" to recommend PCD evaluation, but their justifications may not include discussion of PCD. Indeed, some justifications of a "Yes" PCD evaluation recommendation assert a wholly different diagnosis as the likely underlying etiology, particularly if PCD was not discussed in the encounter note (Appendix A.1). Furthermore, in less severe instances, justifications would highlight symptoms not specific or "classic" for PCD as supportive of the diagnosis, which was a common reason for the "partially sound" label (Table 2). In more severe instances, models cited symptoms altogether unrelated to PCD – such as neuromuscular weakness – as supportive of PCD diagnosis, which was a common reason for the "clinically concerning" label. Though LLMs have previously been shown to respond well to clinical vignettes and have shown initial promise in extracting clinically relevant information from unstructured clinical text (Wu et al., 2024; Shyr et al., 2024; Ríos-Hoyo et al., 2024), our findings suggest that when presented with real-world clinical information in the form of an early encounter note—which may include information both contributory and non-contributory to the ultimate diagnosis, or discuss a patient who is struggling with multiple diagnoses—limitations remain in picking out clinically relevant information pertaining to a specific diagnosis, in this case PCD.

GPT-OSS-120B was a notable exception to this theme, as reflected in its high percentage of clinically sound justifications as compared to other models. It not only had more dependable logical alignment of "Yes" or "No" PCD evaluation recommendation and correlating justification, but also more consistently provided PCD-focused justifications and testing, even if PCD was not discussed in the source note (Table A.1). This finding is particularly impactful in the context of our diagnostic performance evaluation, in which GPT-OSS-120B had the lowest scores across performance metrics. We suspect that, were those PCD evaluation recommendations of "Yes" converted to "No" in cases where the model's justification did not actually imply a suspicion for PCD, performance would have been more similar across models. This underscores the importance of continued deeper manual analysis when assessing LLMs' clinical reasoning abilities, as it may reveal fundamentally flawed reasoning underlying seemingly appropriate recommendations.

**Performance on confirmed PCD cases** was particularly revealing. In this subset, clinically sound explanations dropped to as low as 10.7%, while clinically weak explanations rose to 42.9%. At the same time, incomplete testing recommendations were common, ranging from 39.3% to 60.7%



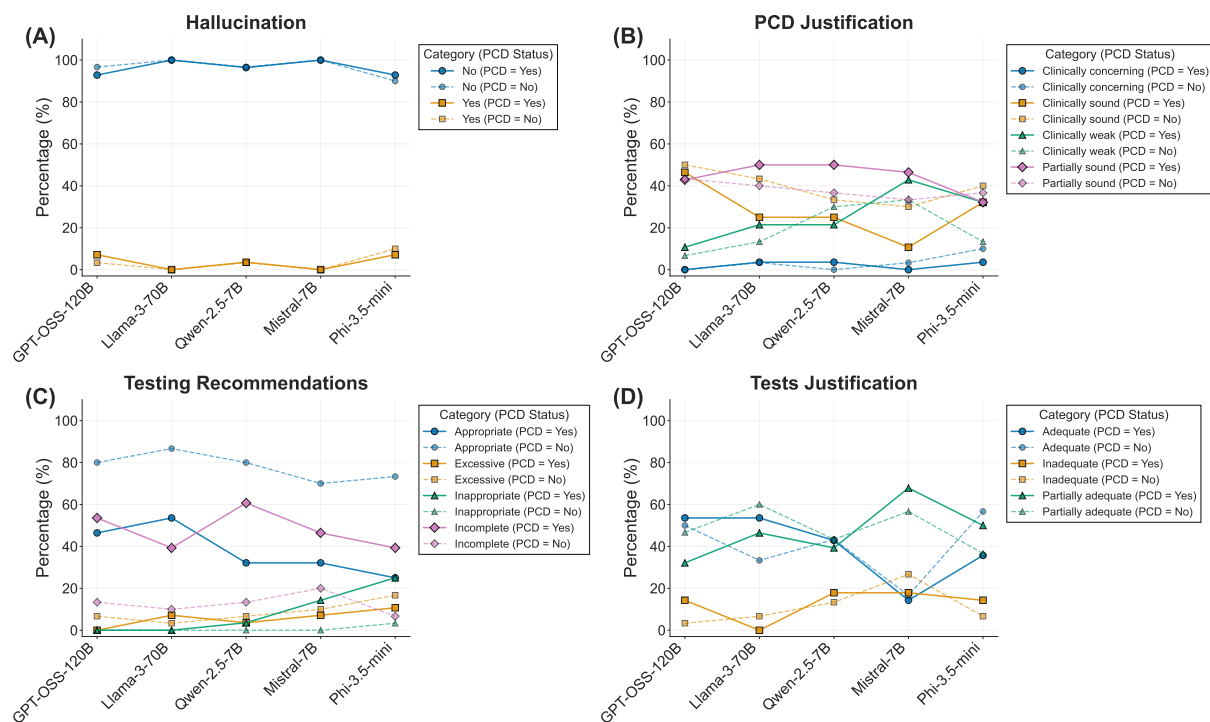


Figure 5: Expert evaluation of five LLMs across four metrics: (A) hallucinations in PCD justifications, (B) strength of clinical reasoning, (C) completeness of testing recommendations, and (D) adequacy of testing justifications. Solid lines indicate PCD-positive cases; dashed lines indicate PCD-negative controls. Colors denote annotation categories. A medical expert systematically evaluated LLM-generated clinical impressions, justifications, and testing recommendations. Overall, models rarely hallucinate (A), but their reasoning and test recommendations are often only partly correct or incomplete (B–D).

across models. This pattern suggests that models struggle most when prioritization and selection of high-yield next steps are required, rather than general symptom summarization. Similar to the observed incompatibility of some PCD evaluation recommendation responses and their justifications, the "incomplete" testing label was often applied to cases where a "Yes" PCD evaluation response was given, but no PCD-specific testing was recommended. In these cases, models frequently recommended workup that was reasonable for ruling out other pulmonary diseases, but uninclusive of important PCD diagnostic tests.

In regard to this discordance, both for PCD evaluation justifications and testing recommendations, one important underlying factor was the source note, as justifications and testing were observed to commonly reflect the note's assessment and plan. Therefore, if PCD was not considered or discussed by the provider in the source note, the model would produce explanations and testing recommendations that aligned more with those etiologies discussed in the note – even when the initial answer for PCD evaluation recommendation was "Yes" (A.1). Even GPT-OSS-120B was impacted, despite its higher percentage of clinically sound, PCD-focused justifications, and was susceptible to

answering "No" to further PCD evaluation in PCD cases where this disease was not discussed in the initial note. This may in part account for its lower sensitivity. Conversely, if the source note contained adequate discussion of PCD, the models were better able to produce clinically sound output. Overall, these findings suggest that the models – particularly when biased by information that already includes proposed diagnoses and testing – are less suited to critically evaluate this input and provide independent conclusions. In the future, prompt adjustment to encourage disagreement with the note, if warranted, may help mitigate some of this issue.

Although inappropriate testing was uncommon overall, it reached 25.0% for Phi-3.5-mini and 14.3% for Mistral-7B in confirmed cases, which could increase clinical burden if outputs are followed without review. Aside from unneeded testing, inappropriate testing labels were applied in instances where the model's recommendation included orders copied from the source note that were not testing, for example, medications or clinic nursing protocols. They were also applied when recommendations failed to consider the patient's age. For example, nasal nitric oxide measurement is a common screening test for PCD, however it cannot be performed in children under 4 years.

| <p>Case x is a former xy week preemie who stayed xy days in NICU. She required CPAP but weaned to RA prior to discharge. Mom states that Case x has daily cough (sometimes to point of projectile vomiting) but since starting Pulmicort in March the cough has improved. She has not needed albuterol in over a month. Case x is a 5m female who is here for initial pulmonary consultation. Since she was born she's had a lot of UAO noise and cough - a lot of mucus production. In the neonatal period she reportedly had RML PNA and RUL atelectasis that prompted chest CT scan to look for CPAM and TEF. She then had presumed RSV infection in early Feb - which is when she started prn albuterol. Her cough worsened so she went to ED, but was better by the time she got to the ED. Pulmicort was started in late xyz month - 0.xx mg BID - still sounds junky and coughs. Cough is better, but still sounds junky and rattly. Last used albuterol over a month ago. She used to have lots of vomiting - only sometimes post tussive. Now just has wet burps. No loud snoring. No gasping. Tested for pertussis because she turned purple with coughing once - was negative. Sometime also happens when she BFs. Mom thinks her let down may be fast. Wheezing on exam has been reported when she was sick. Albuterol helps the cough. Last used albuterol a month. Takes albuterol via neb. Difficult to give her a nebulizer tx - she fights it. No h/o recurrent pneumonia or ear infections. Has never needed systemic steroids. No H/O eczema. No h/o murmur or seizures. Growing well and eating well. Normal stools. I reviewed an ED visit note from xyz month for resp distress not responsive to albuterol at home in which lungs were CTA and labs were reassuring. Tolerated po intake and was sent home to t/u with PCP. Allergies &amp; Medications: .....</p> |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |                     |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|
| LLM                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | LLM Response                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | Expert's Evaluation |
| GPT-OSS-120B                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | PCD: No                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | ×                   |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Justification: The infant's chronic cough and wheeze are explained by mild persistent asthma and upper airway congestion, with documented response to inhaled corticosteroids and albuterol. There is no history of situs inversus, persistent neonatal respiratory failure, chronic sinusitis, or otitis media that would raise suspicion for primary ciliary dyskinesia. The clinical picture aligns with asthma rather than a motile cilia disorder.                                                                                                       | Clinically weak     |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Testing: None                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Incomplete          |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Justification: Given the clear asthma diagnosis and lack of PCD red flags, additional PCD-specific testing (e.g., nasal nitric oxide, ciliary ultrastructure) is not indicated at this time.                                                                                                                                                                                                                                                                                                                                                                  | Inadequate          |
| Llama-3-70B                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | PCD: No                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | ×                   |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Justification: The patient's symptoms and medical history do not suggest primary ciliary dyskinesia (PCD). The patient's cough and wheezing are likely related to mild asthma and upper airway congestion, which are being managed with Pulmicort and albuterol. There is no mention of situs inversus, nasal polyposis, or other characteristic features of PCD.                                                                                                                                                                                             | Partially sound     |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Testing: None                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Incomplete          |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Justification: Given the patient's symptoms and diagnosis of mild asthma, no further testing is recommended at this time. The patient's symptoms are being managed with current medications and therapies, and there is no indication for additional testing.                                                                                                                                                                                                                                                                                                 | Partially adequate  |
| Qwen-2.5-7B                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | PCD: Yes                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | ✓                   |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Justification: The patient has a history of wheezing, cough, and upper airway symptoms, which are suggestive of asthma. However, the patient also has persistent upper airway noise and mucus production, which may indicate underlying airway inflammation or other conditions. A detailed PCD evaluation is recommended to rule out other causes and to optimize her asthma management.                                                                                                                                                                     | Clinically sound    |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Testing: PFTs, chest CT, and possibly allergy testing                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Incomplete          |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Justification: PFTs will help assess the degree of airway obstruction and responsiveness to bronchodilators. Chest CT can provide more detailed imaging to evaluate the extent of airway inflammation and other structural abnormalities. Allergy testing can help identify potential triggers for her symptoms.                                                                                                                                                                                                                                              | Adequate            |
| Mistral-7B                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | PCD: Yes                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | ✓                   |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Justification: The patient has a history of prematurity, prolonged NICU stay, RSV infection, and recurrent cough with mucus production. She has required CPAP and has had a chest CT scan to look for CPAM and TEF. She has also had a history of wheezing and improved with albuterol. The patient's cough persists despite Pulmicort and still sounds junky and rattly.                                                                                                                                                                                     | Partially sound     |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Testing: Pulmonary function tests (PFTs) and echocardiogram                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | Inappropriate       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Justification: To further evaluate the patient's lung function and potential cardiac involvement.                                                                                                                                                                                                                                                                                                                                                                                                                                                             | Inadequate          |
| Phi-3.5-mini                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | PCD: Yes                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | ✓                   |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Justification: The patient has a history of respiratory symptoms including cough, wheezing, and upper airway congestion. Despite being on Pulmicort, symptoms persist, suggesting suboptimal control. The chest CT scan indicates consolidations and ground-glass opacities, which could be indicative of underlying asthma or bronchitis. The patient's growth and dietary changes may also impact reflux, which can exacerbate respiratory symptoms. Given the chronic nature of the symptoms and the findings on imaging, further evaluation is warranted. | Partially sound     |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Testing: Spirometry, allergy testing, and possibly a trial of inhaled corticosteroids (ICS) via MDI for better control.                                                                                                                                                                                                                                                                                                                                                                                                                                       | Inappropriate       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Justification: Spirometry will assess the degree of airflow obstruction and help confirm the diagnosis of asthma. Allergy testing may identify specific triggers contributing to the patient's symptoms. A trial of ICS via MDI could provide better symptom control and adherence compared to nebulizer use, which is currently causing resistance due to the patient's age and difficulty with nebulizer administration.                                                                                                                                    | Adequate            |

Table 2: LLM responses and annotator evaluation for a sample case. The top row shows a partial excerpt of the input clinical note. Appendix A.1 shows another case with *clinically weak* justifications.

**Review of test justifications** further clarify limitations. Adequate justification of recommended tests ranged from 15.5% to 51.7% overall, with partially adequate explanations dominating for several models. In confirmed PCD cases, adequate justification was only 14.3% for Mistral-7B and 35.7% for Phi-3.5-mini, compared with 53.6% for GPT-OSS-120B and Llama-3-70B. This gap matters clinically because decision support tools must explain not only what to do, but why a given action is appropriate for a specific patient. The partially adequate label was commonly applied due to limited rationale given for a specific test. For outputs that did expound more on testing justification, this label was often given due to recurring, subtle inaccuracies observed. For example, models did not have a solid

grasp on which PCD diagnostic tests measure ciliary structure versus ciliary function, and there was a tendency to label a certain test as "necessary" for diagnosis of PCD, when in reality no one test is considered a gold standard test for its diagnosis (Knowles et al., 2016; Nussbaumer et al., 2021).

**Inadequate test justifications**, meanwhile, were overall rare, but did reach as high as 26.7% for Mistral-7B. Inadequate labels were given for vague outputs failing to justify testing in any meaningful way, as well as for misleading reasoning. For the latter, one inconsistent but notable mistake was asserting that tests used to diagnose other diseases are diagnostic of PCD. In particular, *CFTR* genetic testing and/or sweat chloride testing for cystic fibrosis (CF) — a disease with features similar to PCD

— are often pursued in PCD workup to rule out CF. However, these tests were at times either implied or definitively proclaimed to diagnose PCD. This issue was more prevalent with Qwen 2.5 7B and Phi-3.5-mini, which both made the clinically concerning mistake of calling sweat testing the "gold standard for PCD diagnosis." For example, in one confirmed PCD case, GPT-OSS-120B provided the following justification: "*Nasal NO is a sensitive screening tool; genetic testing can identify pathogenic mutations; HSVM and TEM provide functional and structural confirmation of ciliary dysfunction*". This was annotated as 'adequate'. In contrast, for the same case, Phi-3.5-mini stated: "*The sweat chloride test measures the amount of salt in sweat, which is typically elevated in PCD*", which was annotated as inadequate due to inaccurately asserting sweat testing as a test for PCD. Together with the more subtle inaccuracies above, these findings emphasize the need for oversight from an experienced clinician in testing prioritization and interpretation, both to grasp nuances and to avoid fundamental misunderstanding of tests recommended by LLMs.

### 5.3. Clinical Implications

Taken together, these findings suggest that no single model is uniformly best across all dimensions, even across models spanning approximately 7B to 120B parameters. Phi-3.5-mini and the ensemble achieve the highest sensitivity, but Phi-3.5-mini also shows the highest hallucination rate (8.6%) and notable inappropriate testing in confirmed cases. GPT-OSS-120B produces the highest proportion of clinically sound explanations (48.3%) and adequate test justifications (51.7%), with no inappropriate testing, but lower sensitivity. Llama-3-70B shows zero hallucinations and the highest overall rate of appropriate testing (70.7%), but weaker diagnostic explanation quality. Overall, current models appear best suited for early signal amplification and case review rather than independent clinical decision-making. Improving prioritization, justification quality, and consistency across models remains a key challenge for future work.

## 6. Conclusion & Future Work

Using real first-visit pediatric pulmonology notes, this study shows that several openly available language models can spot early signs that should trigger a closer look for primary ciliary dyskinesia, even before the diagnosis is known. Across models, the main strength was catching most true cases, but they also flagged many children who did not have the disease, which could lead to extra referrals or testing if used without clinician judgment. A clinician review showed that even though model outputs

stayed close to the note, their explanations and test choices were often only partly convincing, especially when choosing the most useful next steps that mattered most. In particular, despite high sensitivity when affirmatively recommending additional PCD evaluation, explanations of that recommendation — as well as recommended testing and justification of that testing — often did not appropriately reflect PCD as the disease process in question, which is a concerning weakness if the goal is to use these models as clinical decision making aids. Overall, these systems look more suitable as a cautious "early warning" aid to prompt review, not as a standalone decision maker. Our results justify larger studies across more hospitals and longer patient records to judge real-world usefulness.

## 7. Limitations

While this study demonstrates that LLMs can identify diagnostic signals from routine clinical notes, the findings should be interpreted as an initial step. The sample size was limited to 58 patients from a single pediatric hospital, which may affect generalizability to other settings and populations. We analyzed only the first pulmonology note per patient to reflect early referral, but later clinical information was not included. The models were provided with only one provider note. A task more reflective of clinical practice would be to review multiple provider notes, or even the lifetime accumulated notes and results within an EMR, to guide PCD diagnostic decision making. However, the volume of chart review, input limitations, as well as previous studies suggesting that larger volumes of information may actually impede LLM performance (Tang et al., 2023), made this approach impractical. Given this, we opted to utilize the first encounter note with a pulmonology specialist, as it tends to be a critical juncture where the possibility of disease exists, but diagnostic uncertainty remains — therefore an optimal time where providers or patients might turn to LLMs for guidance on evaluation. The evaluation focused on open source models with a single instruction setup, and model explanations were reviewed by one clinician. Future work should include larger and more diverse datasets, multiple institutions, longitudinal notes, additional models and prompting strategies, and review by multiple clinicians to better assess robustness and clinical usefulness.

## 8. Ethical Considerations

Models were evaluated retrospectively and not used for real clinical decisions. We emphasize that such systems are intended for research on clinical text understanding and should not be used

as standalone diagnostic tools without appropriate clinical oversight. The authors declare no competing interest.

## 9. Bibliographical References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacrose, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Pra-neetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#).
- Tassallah Abdullahi, Ritambhara Singh, and Carsten Eickhoff. 2024. [Learning to Make Rare and Complex Diagnoses With Generative AI Assistance: Qualitative Study of Popular Large Language Models](#). *JMIR Medical Education*, 10(1):e51391. Company: JMIR Medical Education Distributor: JMIR Medical Education Institution: JMIR Medical Education Label: JMIR Medical Education Publisher: JMIR Publications Inc., Toronto, Canada.
- Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *arXiv preprint arXiv:2508.10925*.
- AI@Meta. 2024. [Llama 3 model card](#).
- Laura Behan, Audrey Dunn Galvin, Bruna Rubbo, Sarah Masefield, Fiona Copeland, Michele Manion, Bernhard Rindlisbacher, Beatrice Redfern, and Jane S Lucas. 2016. [Diagnosing primary ciliary dyskinesia: an international patient perspective](#). *European Respiratory Journal*, 48(4):1096–1107.
- Gully Burns, Carey Kauffman, Michele Manion, Ruth-Anne Pai, Carlos Milla, Michael G. O'Connor, Adam J. Shapiro, and Heidi Bjornson-Pennell. 2025. [Feasibility of machine learning analysis for the identification of patients with possible primary ciliary dyskinesia](#). *Orphanet Journal of Rare Diseases*, 20(1):516.
- 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 the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 4850–4861, New York, NY, USA. Association for Computing Machinery.
- Sudeshna Das, Yao Ge, Yuting Guo, Swati Rajwal, JaMor Hairston, Jeanne Powell, Drew Walker, Snigdha Peddireddy, Sahithi Lakamana, Selen Bozkurt, et al. 2025. [Two-layer retrieval-augmented generation framework for low-resource medical question answering using reddit data: proof-of-concept study](#). *Journal of Medical Internet Research*, 27:e66220.
- Karamarie Fecho, Stanley C Ahalt, Michael Knowles, Ashok Krishnamurthy, Margaret Leigh, Kenneth Morton, Emily Pfaff, Max Wang, and Hong Yi. 2022. [Leveraging open electronic health record data and environmental exposures data to derive insights into rare pulmonary disease](#). *Frontiers in Artificial Intelligence*, 5:918888.
- Frank M. Hodges, Sundareswar Pullela, Guy Cohen, Akshay Mulgund, Jared C. Roach, and Stephen A. Ramsey. 2025. [Using AI to Improve Diagnosis and Treatment of Rare Diseases: A Chat Agent for Equitable and Accessible Healthcare](#). In *Artificial Intelligence in*



- Medicine*, pages 186–191, Cham. Springer Nature Switzerland.
- Michael R. Knowles, Maimoona Zariwala, and Margaret Leigh. 2016. [Primary ciliary dyskinesia. \*Clinics in Chest Medicine\*, 37\(3\):449–461. Rare and Orphan Lung Diseases.](#)
- Mistral AI. 2024. [Mistral7b](#). Hugging Face Model. Apache-2.0 License.
- Enock Niyonkuru, J Harry Caufield, Leigh C Carmody, Michael A Gargano, Sabrina Toro, Patricia L Whetzel, Hannah Blau, Mauricio Soto Gomez, Elena Casiraghi, Leonardo Chimirri, Justin T Reese, Giorgio Valentini, Melissa A Haendel, Christopher J Mungall, and Peter N Robinson. 2025. [Leveraging generative AI to assist biocuration of medical actions for rare disease. \*Bioinformatics Advances\*, 5\(1\):vbaf141.](#)
- Mirjam Nussbaumer, Elisabeth Kieninger, Stefan A Tschanz, Sibel T Savas, Carmen Casaulta, Myrofora Goutaki, Sylvain Blanchon, Andreas Jung, Nicolas Regamey, Claudia E Kuehni, et al. 2021. [Diagnosis of primary ciliary dyskinesia: discrepancy according to different algorithms. \*ERJ open research\*, 7\(4\).](#)
- PCD Foundation. 2026. Pcd foundation registry. <https://www.pcdfoundation.org/registry/>. Accessed: 2026-03-29.
- PCD Support UK. 2026. National pcd registry. <https://pcdsupport.org.uk/pcd-registry/>. Accessed: 2026-03-29.
- Swati Rajwal, Eugene Agichtein, and Abeed Sarker. 2024. [EM\\_Mixers at MEDIQA-CORR 2024: Knowledge-enhanced few-shot in-context learning for medical error detection and correction.](#) In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 590–595, Mexico City, Mexico. Association for Computational Linguistics.
- Swati Rajwal, Mary Ellen Fain, Kiri Diaz-Asper, George Lucas Silva, Abeed Sarker, and Lokesh Guglani. 2026. Assessing the performance of large language models for diagnostic reasoning for primary ciliary dyskinesia in pediatric patients. *American Journal of Respiratory and Critical Care Medicine*, 212(Abstract).
- Swati Rajwal, Shivank Garg, Reem Abdel-Salam, and Abdelrahman Zayed. 2025. [Do biased models have biased thoughts?](#) In *Second Conference on Language Modeling*.
- Ann-Katrin Rother, Nicolaus Schwerk, Folke Brinkmann, Frank Klawonn, Werner Lechner, and Lorenz Grigull. 2015. [Diagnostic support for selected paediatric pulmonary diseases using answer-pattern recognition in questionnaires based on combined data mining applications—a monocentric observational pilot study. \*PLoS One\*, 10\(8\):e0135180.](#)
- Alejandro Ríos-Hoyo, Naing Lin Shan, Anran Li, Alexander T. Pearson, Lajos Pusztai, and Frederick M. Howard. 2024. [Evaluation of large language models as a diagnostic aid for complex medical cases. \*Frontiers in Medicine\*, 11. Publisher: Frontiers.](#)
- Alyssa K Salinas, Gerard Smits, Stephen B Spurgin, Tanya M Martínez-Fernandez, and Yadira M Rivera-Sánchez. 2025. [Referral rates and diagnostic evaluation for primary ciliary dyskinesia in patients with laterality defects who meet primary ciliary dyskinesia evaluation criteria. \*Pediatric pulmonology\*, 60\(6\):e71129.](#)
- 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.](#)
- Liyan Tang, Zhaoyi Sun, Betina Ilday, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. [Evaluating large language models on medical evidence summarization. \*NPJ digital medicine\*, 6\(1\):158.](#)
- Osnat Tausky, Eytan Kaplan, Gili Kadmon, Yulia Gendler, Elhanan Nahum, Shai Yitzhaki, and Avichai Weissbach. 2025. [Large language model as a clinical decision support tool in the initial management of critically ill children: a pilot evaluation. \*European Journal of Pediatrics\*, 184\(12\):757.](#)
- Qihong Wei, Zhengxiong Yao, Ying Cui, Bo Wei, Zhezhen Jin, and Ximing Xu. 2024. [Evaluation of chatgpt-generated medical responses: A systematic review and meta-analysis. \*Journal of Biomedical Informatics\*, 151:104620.](#)
- Jinge Wu, Hang Dong, Zexi Li, Haowei Wang, Runci Li, Arijit Patra, Chengliang Dai, Waqar Ali, Phil Scordis, and Honghan Wu. 2024. [A hybrid framework with large language models for rare disease phenotyping. \*BMC Medical Informatics and Decision Making\*, 24\(1\):289.](#)
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei



Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Maimoona A. Zariwala, Michael R. Knowles, and Margaret W. Leigh. 1993–2025. [Primary ciliary dyskinesia](#). In Margaret P. Adam, Ghayda M. Mirzaa, Roberta A. Pagon, Stephanie E. Wallace, Laura J. H. Bean, Kevin W. Gripp, and Anne Amemiya, editors, *GeneReviews®*. University of Washington, Seattle. Updated 2025.

Wangji Zhou, Qiaoling Chen, Yaqi Wang, Anhui Guo, Aohua Wu, Xueqi Liu, Jinrong Dai, Shuzhen Meng, Christopher Situ, Yaping Liu, et al. 2025. [An electronic medical record retrieval system can be used to identify missed diagnosis in patients with primary ciliary dyskinesia](#). *Journal of internal medicine*, 297(1):93–100.

## Appendix

### A. Sample case with LLM output

Table [A.1](#) shows another sample case with predictions and justifications by each LLM. Another case is shown in Table [2](#).

Case X is a XX-month-old female who presents with transfer from XX facility due to concern for possible tracheal deviation, persistent oxygen requirement, and poor oral feeding. Consult requested for home oxygen and outpatient pulmonary follow-up. She is on XX LPM oxygen and being treated for BPD with inhaled corticosteroid (XX dose, XX puffs BID) and albuterol every XX hours. She is also on diuretics (XX). Mild reflux diagnosed on pH probe. Currently on reflux medications (XX) and receiving NG feeds. No desaturation episodes reported. Per review of outside records, she was extubated to CPAP at XX days of life and remained on CPAP for approximately XX duration before transition to high-flow nasal cannula, now on low-flow nasal cannula. She previously received diuretic therapy for approximately XX duration, then transitioned to current regimen prior to transfer. History of pulmonary hypertension treated with inhaled nitric oxide for XX days, now resolved. Continues on low-flow oxygen. Transferred for evaluation of suspected tracheal deviation to rule out airway abnormality.

| LLM          | LLM Response                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | Expert's Evaluation   |
|--------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| GPT-OSS-120B | PCD: No                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | ×                     |
|              | Justification: The infant has chronic lung disease secondary to prematurity (BPD) with a mild right tracheal deviation likely due to a prominent thymus. There is no history of recurrent sinusitis, otitis media, or bronchiectasis, and no laterality defects or unexplained persistent cough that would suggest primary ciliary dyskinesia. The clinical picture is fully explained by prematurity, BPD, and gastroesophageal reflux, making PCD evaluation unnecessary at this time. | Clinically sound      |
|              | Testing: None                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | Incomplete            |
|              | Justification: No additional diagnostic work-up for PCD is indicated because the patient's symptoms are attributable to BPD and prematurity. Routine follow-up imaging (repeat echocardiogram) and clinical monitoring are already planned, and there are no findings that would warrant nasal nitric oxide measurement, ciliary biopsy, or genetic testing for PCD.                                                                                                                     | Adequate              |
| Llama-3-70B  | PCD: Yes                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | ✓                     |
|              | Justification: The patient has a history of bronchopulmonary dysplasia (BPD) and is on oxygen therapy, which increases the risk of primary ciliary dyskinesia (PCD). Additionally, the patient has a mild tracheal deviation, which could be related to PCD. Further evaluation is necessary to rule out PCD.                                                                                                                                                                            | Clinically concerning |
|              | Testing: Nasal nitric oxide measurement, high-speed video microscopy analysis of cilia, and genetic testing                                                                                                                                                                                                                                                                                                                                                                              | Appropriate           |
|              | Justification: These tests are recommended to evaluate the patient's ciliary function and structure, which can help diagnose PCD. Nasal nitric oxide measurement is a non-invasive test that can indicate PCD, while high-speed video microscopy analysis of cilia can assess ciliary beat frequency and pattern. Genetic testing can identify mutations associated with PCD.                                                                                                            | Adequate              |
| Qwen-2.5-7B  | PCD: Uncertain                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | ×                     |
|              | Justification: The patient has a history of BPD, mild tracheal deviation, and potential pulmonary hypertension. While the tracheal deviation could be secondary to a prominent thymus, further evaluation is warranted to rule out other causes and to monitor the patient's condition.                                                                                                                                                                                                  | Clinically weak       |
|              | Testing: CT of the chest with contrast                                                                                                                                                                                                                                                                                                                                                                                                                                                   | Incomplete            |
|              | Justification: To better evaluate the tracheal deviation and rule out other causes, a CT of the chest with contrast is recommended. This will provide more detailed imaging to assess the trachea and other structures in the chest.                                                                                                                                                                                                                                                     | Adequate              |
| Mistral-7B   | PCD: Yes                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | ✓                     |
|              | Justification: The patient has a history of BPD, is on oxygen, and has a mildly dilated RV, thickened RV wall, and a probable left coronary cameral fistula to the RV on echocardiogram, suggesting possible pulmonary hypertension.                                                                                                                                                                                                                                                     | Clinically weak       |
|              | Testing: Right Heart Catheterization                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Excessive             |
|              | Justification: To confirm the diagnosis of pulmonary hypertension and guide treatment                                                                                                                                                                                                                                                                                                                                                                                                    | Adequate              |
| Phi-3.5-mini | PCD: Yes                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | ✓                     |
|              | Justification: The patient has a history of BPD, mild tracheal deviation, and a probable coronary fistula, which could contribute to respiratory symptoms. The presence of chronic lung disease and the need to rule out pulmonary hypertension further support the need for PCD evaluation.                                                                                                                                                                                             | Clinically weak       |
|              | Testing: Echocardiogram with contrast, repeat chest CT, and possibly a cardiac MRI.                                                                                                                                                                                                                                                                                                                                                                                                      | Excessive             |
|              | Justification: The echocardiogram with contrast is recommended to further evaluate the suspected coronary fistula and to rule out pulmonary hypertension. A repeat chest CT is suggested to monitor the chronic lung disease and tracheal deviation. A cardiac MRI may provide additional information on the coronary fistula and cardiac structure.                                                                                                                                     | Adequate              |

Table A.1: LLM responses and annotator evaluation for a sample case. The top row shows a partial excerpt of the input clinical note.

# Beyond One-Size-Fits-All: Multi-Agent Refinement Framework for Persona-Based Biomedical Summarization

Rohan Charudatt Salvi<sup>1</sup>, Chirag Chawla<sup>2</sup>, Md. Shad Akhtar<sup>3</sup>, Shweta Yadav<sup>1</sup>

<sup>1</sup>University of Illinois Chicago <sup>2</sup>Indian Institute of Technology Varanasi

<sup>3</sup>Indraprastha Institute of Information Technology Delhi

{rcsalvi2, shwetay}@uic.edu

## Abstract

Biomedical lay summarization aims to make biomedical research accessible to non-experts, but most approaches assume a uniform audience, overlooking variation in medical literacy and information needs. We present MAPS (Multi-Agent Persona-based Summarization), a framework that generates persona-specific summaries through iterative cross-agent feedback. Human evaluation shows MAPS improves quality over single-agent baselines, while automatic metrics fail to capture these gains. LLM-based judges also exhibit limited sensitivity, assigning inflated scores and misdetecting errors. These findings highlight the need for improved evaluation methods for persona-based summarization.

**Keywords:** Biomedical Text Summarization, Controlled Text Generation, Multi-Agent System, Summary Evaluation

## 1. Introduction

Lay summarization aims to summarize jargon-heavy texts for non-experts in plain language (King et al., 2017). The lay summaries make scientific knowledge accessible beyond specialized communities (Stableford and Mettger, 2007). Prior lay summarization work has primarily targeted a uniform non-expert audience (Guo et al., 2021; Goldsack et al., 2023). In the real world, however, information needs vary widely across audiences. A pre-medical student benefits from more technical depth than a lay person, while a researcher looks for methodological detail that neither would expect (Salvi et al., 2025a). These differences motivate *persona-based summarization*, which requires controlling not only linguistic complexity but also content selection, such as what to explain, retain, and how much detail to include.

Prior controllable summarization work has focused on linguistic attributes such as sentence length or reading level (Luo et al., 2022; Tran et al., 2025b), without modeling how different audiences require different content, not just different wording. The PERCS dataset (Salvi et al., 2025a) addresses this by providing summaries tailored to four personas that vary in both readability and information depth, but effective generation methods for producing such persona-specific summaries remain underexplored.

Persona-based summarization requires balancing competing objectives such as readability, faithfulness, and persona-appropriate detail that are difficult to jointly optimize (Mullick et al., 2024; Song et al., 2025). Multi-agent LLM frameworks offer a natural solution by decomposing generation into specialized subtasks (Fang et al., 2025; Mo and Hu, 2024), and have shown promise for text sim-

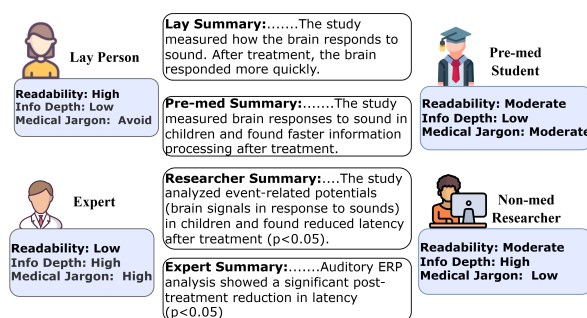


Figure 1: Persona-specific profiles highlighting differences in objectives and communication needs.

plification (Fang et al., 2025) and lay summarization (Lyu and Pergola, 2024). However, applying multi-agent LLMs to persona-based summarization raises design questions. Feedback must reflect persona-specific criteria, since what counts as an error for a lay reader (e.g., unexplained jargon) differs from what matters for a researcher (e.g., omitted statistics). Adjudication must similarly balance readability and detail differently per persona.

To address these challenges, this paper presents **MAPS** (Multi-Agent Persona-based Summarization), a framework for generating persona-specific biomedical summaries through iterative cross-agent feedback. Using PERCS, we study how design choices affect persona alignment by comparing feedback strategies and adjudication mechanisms. We also examine whether automatic metrics and LLM-based judges reliably evaluate persona-specific summaries. Our results show MAPS improves summary quality in human evaluation, while automatic metrics and LLM judges exhibit weak alignment with human judgments, highlighting a key evaluation challenge.

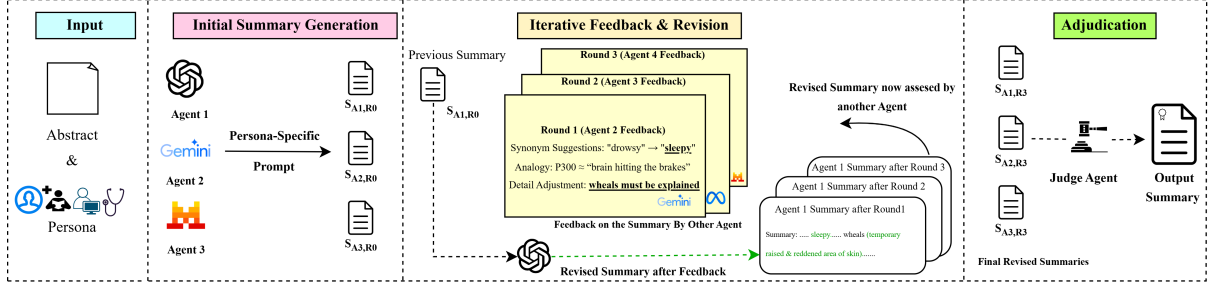


Figure 2: Overview of the MAPS pipeline ( $S_{i,j}$ : summary from agent  $i$  in feedback round  $j$ ).

## 2. Background

Prior work on biomedical lay summarization has explored in-context learning and prompting-based methods (Goldsack et al., 2025; Ming et al., 2025), largely focusing on controlling linguistic attributes such as reading level (Luo et al., 2022; Tran et al., 2025a). Persona-dependent content selection remains underexplored. The PERCS dataset (Salvi et al., 2025a) addresses this gap by enabling persona-based evaluation and establishing single-LLM baselines, which we extend by studying multi-agent generation. Multi-agent LLM frameworks decompose generation into specialized subtasks to better address various objectives such as faithfulness, readability, and attributes like length (Fang et al., 2025)(Fang et al., 2025; Mo and Hu, 2024; Lyu and Pergola, 2024), though prior multi-agent work reports mixed results on automatic metrics (Zhu et al., 2025). In contrast, we evaluate summaries using human evaluation, LLM judges, and automatic metrics to better assess persona alignment.

## 3. Method

### 3.1. MAPS Framework

Given a biomedical abstract  $D$  and target persona  $P$ , MAPS generates a summary  $S$  that is faithful to  $D$  while matching persona-specific readability and information depth. MAPS uses  $k$  agents  $\mathcal{A} = \{A_1, \dots, A_k\}$  to generate, iteratively refine, and adjudicate persona-specific summaries (Figure 2).

**Stage 1: Initial Generation** Each agent generates a persona-conditioned summary, yielding diverse candidates  $\{S_i^{(0)}\}$ .

**Stage 2: Iterative Feedback & Revision** Initial summaries contain factual errors, omit information, or exhibit persona mismatches (Fang et al., 2024). MAPS refines them through  $T$  rounds of cross-agent feedback: in each round, agent  $A_i$  receives critique from agent  $A_{(i \bmod k)+1}$  in round-robin scheme and revises its summary. Separating

critique from generation reduces error reinforcement common in self-refinement (Wadhwa et al., 2024; Xie et al., 2025). We explore two feedback strategies:

**Persona-Centered Feedback:** This strategy explicitly specifies expectations for each target persona in terms of content depth, readability level, assumed biomedical knowledge, and summary length. Rather than leaving agents to infer audience needs implicitly, we define persona-specific feedback categories covering simplification requirements, level of detail, key point coverage, and faithfulness. For instance, feedback for a researcher persona emphasizes methodological and numerical details, whereas feedback for a lay persona prioritizes accessible definitions and reduces complexity. Agents evaluate summaries against these criteria and suggest specific revisions to improve persona alignment and faithfulness. As shown in Figure 3, persona-centered feedback highlights a missing definition of "ERP" and notes omitted p-values essential for a research audience. Based on prior work (You et al., 2024; Ming et al., 2025), we also developed a variant *Persona-Centered-Wiki* that provides Wikipedia as a tool to feedback agents, enabling them to verify definitions for faithfulness.

**Error-Structured Feedback:** Rather than aligning summaries through explicit persona profiles,

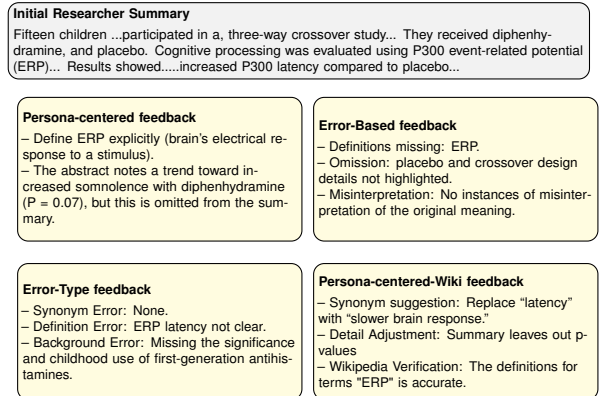


Figure 3: Illustrative examples of feedback strategies employed in the MAPS framework.

this strategy draws on research documenting common errors in lay summaries (Guo et al., 2024; Joseph et al., 2024). The feedback agent assumes the role of a medical expert reviewing summaries against a structured error taxonomy. Unlike persona-centered feedback, this approach is diagnostic rather than prescriptive: agents flag errors without suggesting corrections, leaving revision to the generating agent. We adopt the PERCS taxonomy (Salvi et al., 2025a), comprising 11 error types across four categories: new information errors (incorrect definitions, synonyms, background), inference errors (contradictions, omissions, misinterpretations), structure and readability issues, and hallucinations. We test two variants: *Error-based* flags all categories simultaneously, while *Error-type* targets one category per round. As shown in Figure 3, this approach identifies similar issues as persona-centered feedback but categorizes them systematically (e.g., "Definition Error. ERP latency not fully clear").

**Stage 3: Adjudication** After  $T$  rounds, an adjudication step selects a summary among candidates based on comprehensiveness, readability, and faithfulness. We explore three strategies: (a) **Judge Agent**, where a dedicated agent selects the best candidate, (b) **Voting**, where agents vote independently with ties resolved by the judge, and (c) **Combined Generation**, where a judge synthesizes a new summary by merging strengths across all candidates.

## 3.2. Experimental Setup

We evaluate on the PERCS dataset (Salvi et al., 2025a), which contains 500 biomedical abstracts paired with 2,000 expert-curated summaries across four personas: layperson, pre-medical student, non-medical researcher, and medical expert. We use the 150-abstract test split with  $k=4$  agents and  $T=3$  feedback rounds, comparing against zero-shot, few-shot, and self-refine baselines using the same LLMs as PERCS (GPT-4o, Mixtral-8x7B-Instruct, Gemini-2.0 Flash Lite, and LLaMA-3 70B). For MAPS, we test single-model (agents from the same LLM) with persona-based feedback only, and multi-model (agents from diverse LLMs) settings across all feedback strategies and adjudication methods. We include full prompt templates, agent instructions, and evaluation guidelines in the appendix, where we provide the Researcher persona as a representative example. Prompts for all other personas, along with the full codebase, are publicly available at [GitHub](#). A token consumption and cost comparison between baselines and MAPS is provided in Table 13 in the Appendix.

## 3.3. Evaluation

**Automatic Metrics.** We use ROUGE-1,2,L (Lin, 2004) and SARI (Xu et al., 2016) for comprehensiveness, and FKGL, DCRS, CLI, and LENS (Maddala et al., 2022) for readability.

**Human and LLM Evaluation.** Both human raters and LLM evaluate summaries across criteria following Salvi et al. (2025b) guidelines on a 5-point Likert scale: i) *Comprehensiveness*: the extent to which the summary contains information necessary for the target persona to understand the research, ii) *Readability*: how easy the summary is to read based on appropriate medical jargon simplification. iii) *Factuality*: the extent to which the summary is factually consistent with the source abstract.

For human evaluation, we recruited two raters per persona via Upwork, selecting annotators whose background aligned with the target persona. Each annotator rated summaries from three methods (GPT-4o zero-shot as a consistent baseline across all personas, the best-performing PERCS model per persona, and MAPS) for 30 abstracts in their matched persona. To ensure consistent annotation, we measured inter-rater agreement using Krippendorff’s  $\alpha$ , which exceeded 0.84 for all personas. For LLM-based rating, we prompt LLaMA-3 with persona definitions and criteria to rate the same summaries. Additionally, we evaluate whether LLMs can reliably identify errors by prompting LLaMA-3, Mixtral, and Gemini to classify errors according to the PERCS taxonomy on 50 summaries and comparing against medical expert judgments. Notably, we selected LLaMA-3 as the LLM judge based on error detection performance, as it showed the closest alignment with human judgments, particularly for omission errors.

## 4. Results

### (1) Does MAPS Setup Improves Performance?

We compare MAPS (persona-specific feedback with combination adjudication) against GPT-4o zero-shot and the best-performing PERCS model per persona. Table 1 reports results for the selected configurations. On automatic metrics, MAPS does not consistently outperform baselines. However, human evaluation shows a contrasting trend where it substantially outperforms both PERCS models and GPT-4o for Pre-med, Researcher, and Expert personas. For example, MAPS improves Researcher comprehensiveness from 3.36 (Mixtral) to 4.74, and Expert comprehensiveness from 4.33 (GPT-4o) to 4.85. For Layperson, MAPS performs comparably to GPT-4o. Overall, multi-agent feedback markedly improves persona alignment for advanced personas while yield-



| Persona    | Method Category | Method             | R-1 ↑  | R-2 ↑  | R-L ↑  | SARI ↑  | FKGL  | DCRS  | CLI   | LENS    |
|------------|-----------------|--------------------|--------|--------|--------|---------|-------|-------|-------|---------|
| Layperson  | Single LLM      | GPT-4o Few-shot    | 0.6016 | 0.2476 | 0.3554 | 53.5134 | 11.22 | 8.94  | 11.99 | 75.0276 |
|            |                 | GPT-4o Zero-shot   | 0.5972 | 0.2439 | 0.3390 | 53.0256 | 10.37 | 8.54  | 11.30 | 51.7573 |
|            | MAPS            | Combine            | 0.5478 | 0.2067 | 0.2914 | 50.6332 | 12.80 | 11.54 | 13.60 | 75.6658 |
| Premed     | Single LLM      | Llama-3.1 Few-shot | 0.6149 | 0.2926 | 0.3823 | 51.9053 | 16.72 | 11.78 | 17.69 | 53.6619 |
|            |                 | GPT-4o Zero-shot   | 0.5916 | 0.2380 | 0.3434 | 49.5828 | 17.91 | 12.77 | 19.44 | 46.2652 |
|            | MAPS            | Combine            | 0.5962 | 0.2657 | 0.3469 | 50.7323 | 14.96 | 10.35 | 15.72 | 50.1168 |
| Researcher | Single LLM      | Mixtral Few-shot   | 0.6122 | 0.3072 | 0.3978 | 49.4988 | 15.05 | 10.65 | 15.93 | 59.3144 |
|            |                 | GPT-4o Zero-shot   | 0.6207 | 0.2742 | 0.3831 | 43.9918 | 15.87 | 11.00 | 16.48 | 51.2367 |
|            | MAPS            | Combine            | 0.6106 | 0.2815 | 0.3677 | 49.8121 | 16.21 | 11.01 | 16.99 | 52.0884 |
| Expert     | Single LLM      | GPT-4o Zero-shot   | 0.6000 | 0.2905 | 0.3997 | 42.1944 | 17.91 | 12.77 | 19.44 | 51.2367 |
|            | MAPS            | Combine            | 0.6148 | 0.3576 | 0.4213 | 47.8559 | 17.24 | 11.98 | 18.26 | 52.0884 |

Table 1: Selected performance on PERCS across personas using automatic evaluation metrics.

| Lay |           |             |             |             | Pre-med |           |             |             |             |
|-----|-----------|-------------|-------------|-------------|---------|-----------|-------------|-------------|-------------|
| J   | Method    | C           | R           | F           | J       | Method    | C           | R           | F           |
| H   | GPT-4o-ZS | 4.83        | <b>4.97</b> | 4.87        | H       | GPT-4o-ZS | 4.65        | 4.45        | 4.93        |
| H   | GPT-4o-FS | <b>4.87</b> | <b>4.97</b> | 4.93        | H       | Llama3-FS | 4.52        | 4.12        | 4.8         |
| H   | MAPS      | 4.83        | 4.93        | 4.83        | H       | MAPS      | <b>4.81</b> | <b>4.71</b> | <b>5.00</b> |
| L   | GPT-4o-ZS | <b>4.83</b> | 4.57        | 4.94        | L       | GPT-4o-ZS | <b>4.96</b> | 4.75        | <b>4.97</b> |
| L   | GPT-4o-FS | <b>4.83</b> | 4.53        | <b>4.96</b> | L       | Llama3-FS | 4.6         | 4.8         | 4.96        |
| L   | MAPS      | 4.71        | <b>4.63</b> | <b>4.96</b> | L       | MAPS      | 4.86        | <b>4.94</b> | 4.96        |

| Researcher |            |             |             |             | Expert |           |             |             |             |
|------------|------------|-------------|-------------|-------------|--------|-----------|-------------|-------------|-------------|
| J          | Method     | C           | R           | F           | J      | Method    | C           | R           | F           |
| H          | GPT-4o-ZS  | 4.69        | <b>4.5</b>  | <b>4.93</b> | H      | GPT-4o-ZS | 4.33        | <b>4.06</b> | 4.36        |
| H          | Mixtral-FS | 3.36        | 3.21        | 3.71        | H      | MAPS      | <b>4.85</b> | 4.15        | <b>4.91</b> |
| H          | MAPS       | <b>4.74</b> | 4.47        | <b>4.93</b> | L      | GPT-4o-ZS | <b>4.99</b> | <b>5.00</b> | <b>4.99</b> |
| L          | GPT-4o-ZS  | <b>4.30</b> | <b>5.00</b> | <b>4.97</b> | L      | MAPS      | 4.90        | <b>5.00</b> | 4.90        |
| L          | Mixtral-FS | 4.1         | 4.83        | 4.77        |        |           |             |             |             |
| L          | MAPS       | 4.18        | <b>5.00</b> | <b>4.97</b> |        |           |             |             |             |

Table 2: LLM and human evaluation scores across methods and personas measured by Comprehensiveness (C), Readability (R), and Faithfulness (F), with Judge (J) indicating Human (H) or LLM (L).

ing comparable performance for lay summarization.

**(2) Are Automatic Metrics Suitable?** Our results show automatic metrics fail to capture persona-appropriate quality.

**Comprehensiveness:** Weak ROUGE-human alignment is documented in biomedical summarization (Wang et al., 2023). Our human evaluation further shows that information depth varies across personas, a distinction human raters capture but lexical metrics do not. Mixtral Few-shot achieves the best ROUGE for Researcher yet receives drastically lower human Comp. (3.36 vs. 4.74 for MAPS), indicating lexical overlap does not capture persona-appropriate information. SARI, designed for simplification, is misaligned with expert personas where preserving terminology is desirable, decreasing monotonically from Layperson (53.51) to Expert (42.19) and penalizing summaries that retain medical language.

**Readability:** Lower readability scores align with lay personas and higher scores with experts, but the metrics fail to distinguish intermediate personas. Pre-med (14.96-17.91) and Researcher (15.05-16.21) overlap substantially and differ in information depth while using similar levels of jargon (medical and scientific respectively). LENS, trained

| Error Type             | Gemini | Llama | Mixtral | Experts |
|------------------------|--------|-------|---------|---------|
| Background Information | 0      | 6     | 3       | 0       |
| Definition             | 2      | 1     | 3       | 0       |
| Synonym                | 0      | 1     | 0       | 0       |
| Entity                 | 4      | 5     | 7       | 0       |
| Contradiction          | 0      | 0     | 0       | 0       |
| Misinterpretation      | 1      | 2     | 6       | 3       |
| Omission               | 2      | 26    | 36      | 74      |
| Jumping to Conclusions | 0      | 2     | 0       | 1       |
| Structure              | 0      | 0     | 2       | 0       |
| Persona Relevance      | 0      | 2     | 2       | 0       |
| Hallucination          | 0      | 0     | 0       | 2       |

Table 3: Comparison between LLMs as judges and human experts for error detection on a set of 50 summaries

to predict simplification quality, exhibits similar limitations, inversely correlating with human judgments for non-lay personas. For Pre-med, LLaMA-3.1 achieves higher LENS (53.66) than MAPS (50.12), yet humans rate MAPS substantially better (Comp. 4.81 vs. 4.52, Read. 4.71 vs. 4.12). This highlights that simplification-oriented metrics actively misrank summaries when simplification is not the sole objective.

**(3) How Reliable are LLM-Based Judges?** LLM-based judges diverge substantially from human evaluations. For summary rating (Table 2), LLMs assign higher scores with limited variance and frequently favor baselines over MAPS, even when humans strongly prefer MAPS. Strong annotator agreement across personas ( $\alpha > 0.84$ ) suggests this divergence reflects LLM limitations in capturing persona requirements rather than subjective persona definitions. Similarly, for error detection (Table 3), LLMs substantially underdetect omissions (Human: 74, LLaMA: 26, Mixtral: 36, Gemini: 2) while flagging unsupported errors and missing hallucinations. These results indicate the need for improved LLM-based evaluation techniques tailored to persona-specific summarization.

**(4) Which Multi-Agent Setup Performs Best?** MAPS with persona-centered prescriptive feedback and combination adjudication consistently performs best across personas. Full results across all configurations are reported in Tables 6-13 in the appendix. We make several observations. First,

multi-model setups outperform single-model configurations, particularly for the Researcher and Expert personas, suggesting that diversity across agents produces more balanced and persona-appropriate summaries. Second, among adjudication strategies, combination adjudication outperforms both judge-only and voting approaches, indicating that synthesizing strengths across candidate summaries is more effective than selecting a single best one.

Third, persona-centered feedback outperforms error-structured variants, which we attribute to its prescriptive nature. Rather than simply flagging problems, it provides agents with concrete and actionable revision targets that are more consistently incorporated. Finally, comparing MAPS against the self-refine baseline shows that cross-agent feedback outperforms single-agent iterative refinement, especially on comprehensiveness for researcher and expert personas, highlighting that multi-agent collaboration drives gains beyond what self-correction alone can achieve. Together, these findings also explain why error-structured feedback is less effective, as LLM judges have limited sensitivity to fine-grained errors, making diagnostic feedback a weaker signal than explicit persona-centered guidance.

## 5. Conclusion

We introduced MAPS, a multi-agent framework for persona-based biomedical summarization. Human evaluation shows MAPS improves summary quality, but current evaluation methods fail to capture these gains. Unlike lay summarization where lower readability signals quality, persona-based summaries require varying objectives, making standard metrics unreliable. LLM judges struggle, assigning higher scores and missing errors like omissions. These findings highlight the need for evaluation frameworks that model persona expectations.

## 6. Limitations

Our work has several limitations that suggest directions for future research. First, we evaluate only on the PERCS dataset, which includes four predefined biomedical personas. While these personas capture meaningful variation in medical literacy, real users may exhibit hybrid or more fine-grained needs. Extending this study to more granular personas, additional domains such as law, climate science, or multilingual settings would help assess its generalizability. Second, although we show that automatic metrics and LLM judges fail to capture persona-specific quality, we do not introduce new evaluation methods to address this gap. Developing persona-aware metrics and reference-free

evaluation approaches remains an important direction for future work. Moreover, our multi-agent framework also considers only round-robin feedback. Exploring alternative feedback structures, such as hierarchical or fully connected schemes, may offer different trade-offs between feedback diversity and coherence. Lastly, our study centers on prompting-based methods, which are effective in this setting, while the exploration of persona-based fine-tuning and reinforcement learning is left to future work.

## 7. Bibliographical References

- Biaoyan Fang, Xiang Dai, and Sarvnaz Karimi. 2024. [Understanding faithfulness and reasoning of large language models on plain biomedical summaries](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9890–9911, Miami, Florida, USA. Association for Computational Linguistics.
- Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. Collaborative document simplification using multi-agent systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912.
- Tomas Goldsack, Carolina Scarton, and Chenghua Lin. 2025. Leveraging large language models for zero-shot lay summarisation in biomedicine and beyond. *arXiv preprint arXiv:2501.05224*.
- Tomas Goldsack, Zhihao Zhang, Chen Tang, Carolina Scarton, and Chenghua Lin. 2023. [Enhancing biomedical lay summarisation with external knowledge graphs](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8016–8032, Singapore. Association for Computational Linguistics.
- 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 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9194–9211, Miami, Florida, USA. Association for Computational Linguistics.
- Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.
- Sebastian Joseph, Lily Chen, Jan Trienes, Hannah Göke, Monika Coers, Wei Xu, Byron Wallace,

- and Junyi Jessy Li. 2024. [FactPICO: Factuality evaluation for plain language summarization of medical evidence](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8437–8464, Bangkok, Thailand. Association for Computational Linguistics.
- Stuart RF King, Emma Pewsey, and Sarah Shailes. 2017. An inside guide to elife digests. *ELife*, 6:e25410.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. [Readability controllable biomedical document summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Chen Lyu and Gabriele Pergola. 2024. [Society of medical simplifiers](#). In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 61–68, Miami, Florida, USA. Association for Computational Linguistics.
- Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2022. Lens: A learnable evaluation metric for text simplification. *arXiv preprint arXiv:2212.09739*.
- Shufan Ming, Yue Guo, and Halil Kilicoglu. 2025. [Towards knowledge-guided biomedical lay summarization using large language models](#). In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 285–297, Albuquerque, New Mexico. Association for Computational Linguistics.
- Kaijie Mo and Renfen Hu. 2024. Expertease: A multi-agent framework for grade-specific document simplification with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9080–9099.
- Ankan Mullick, Sombit Bose, Rounak Saha, Ayan Bhowmick, Pawan Goyal, Niloy Ganguly, Prasennjit Dey, and Ravi Kokku. 2024. [On the persona-based summarization of domain-specific documents](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14291–14307, Bangkok, Thailand. Association for Computational Linguistics.
- Rohan Charudatt Salvi, Chirag Chawla, Dhruv Jain, Swapnil Panigrahi, Md Shad Akhtar, and Shweta Yadav. 2025a. Percs: Persona-guided controllable biomedical summarization dataset. *arXiv preprint arXiv:2512.03340*.
- Rohan Charudatt Salvi, Swapnil Panigrahi, Dhruv Jain, Shweta Yadav, and Md. Shad Akhtar. 2025b. [Towards understanding LLM-generated biomedical lay summaries](#). In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 260–268, Albuquerque, New Mexico. Association for Computational Linguistics.
- Junjie Song, Yiwen Liu, Dapeng Li, Yin Sun, Shukun Fu, Siqi Chen, and Yuji Cao. 2025. Balancing rewards in text summarization: Multi-objective reinforcement learning via hypervolume optimization. *arXiv preprint arXiv:2510.19325*.
- Sue Stableford and Wendy Mettger. 2007. Plain language: a strategic response to the health literacy challenge. *Journal of public health policy*, 28(1):71–93.
- Hieu Tran, Zonghai Yao, Won Seok Jang, Sharmin Sultana, Allen Chang, Yuan Zhang, and Hong Yu. 2025a. Medreadctrl: Personalizing medical text generation with readability-controlled instruction learning. *arXiv preprint arXiv:2507.07419*.
- Hieu Tran, Zonghai Yao, Lingxi Li, and Hong Yu. 2025b. [ReadCtrl: Personalizing text generation with readability-controlled instruction learning](#). In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 19–36, Albuquerque, New Mexico, US. Association for Computational Linguistics.
- Manya Wadhwa, Xinyu Zhao, Junyi Jessy Li, and Greg Durrett. 2024. [Learning to refine with fine-grained natural language feedback](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12281–12308, Miami, Florida, USA. Association for Computational Linguistics.
- Lucy Lu Wang, Julia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey Kuehl, Erin Bransom, and Byron C Wallace. 2023. Automated metrics for medical multi-document summarization disagree with human evaluations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9871–9889.
- Zhihui Xie, Jie Chen, Liyu Chen, Weichao Mao, Jingjing Xu, and Lingpeng Kong. 2025. Teaching language models to critique via reinforcement learning. *arXiv preprint arXiv:2502.03492*.

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.

Zhiwen You, Shruthan Radhakrishna, Shufan Ming, and Halil Kilicoglu. 2024. [UIUC\\_BioNLP at BioLaySumm: An extract-then-summarize approach augmented with Wikipedia knowledge for biomedical lay summarization](#). In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 132–143, Bangkok, Thailand. Association for Computational Linguistics.

Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. 2025. Medagentboard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks. *arXiv preprint arXiv:2505.12371*.

## A. Human Evaluation Guidelines

**Goal.** The goal of this evaluation is to assess the quality of model-generated summaries. Evaluators will rate summaries based on their **comprehensiveness**, **layness**, **usefulness**, and **factuality**.

### Materials Provided for Evaluation.

- Abstract of the article
- Model-generated plain-text summary for the abstract

### Evaluation Criteria.

- We use a 1-5 Likert scale where 1 indicates poor quality and 5 indicates excellent quality.
- The interpretation of each rating for each evaluation facet is defined below.

**Comprehensiveness.** This criterion evaluates how well the summary contains information necessary for a non-expert reader to understand the high-level topic and the significance of the research.

1. The summary is incomplete; an evaluator cannot understand the topic or the significance of the research.
2. The summary is partially complete; an evaluator gains a vague idea of the topic but cannot grasp the significance due to missing key details.
3. The summary allows an evaluator to understand the topic but lacks important details that convey the research's significance.

4. The summary enables an evaluator to understand both the topic and significance, missing only minor details that could enhance understanding.
5. The summary thoroughly covers all necessary information, allowing an evaluator to fully understand the topic and the significance of the research.

**Readability.** Readability is measured based on the use of medical jargon, sentence structure, and whether explanations are provided for technical terms.

1. There is little difference between the plain-text summary and the original abstract.
2. The summary omits a few jargon-heavy sentences or removes some technical words. It is easier to read but does not meaningfully simplify the content.
3. The summary contains a mix of jargon and simple terms, simple and complex sentences, and some definitions. Lay readers may understand the main points but may find certain terms or sentences confusing.
4. The summary is generally easy to understand, with occasional complex sentences or unexplained medical terms.
5. The summary avoids jargon or replaces it with simple synonyms. When this is not possible, it provides sufficient context or explanations. Sentences are simple and clear, making the summary accessible to a general audience.

**Factuality.** Factuality measures the extent to which the plain-text summary is consistent with the information presented in the abstract.

1. The summary alters the findings or methodology, misrepresenting the study.
2. The summary alters parts of the study in ways that may lead to misinterpretation of the methods or results.
3. The summary is largely accurate but contains frequent minor inconsistencies, such as incorrect figures, typos, or omission of key findings.
4. The summary is accurate with one or two minor exceptions.
5. The summary is fully factual and aligns completely with the abstract.

### Procedure.

| Error                  | Description                                                           |
|------------------------|-----------------------------------------------------------------------|
| Incorrect Definitions  | Wrong or misleading explanations of medical terms or concepts.        |
| Incorrect Synonyms     | Replacing medical words with inaccurate or oversimplified terms.      |
| Incorrect Background   | False or irrelevant contextual details about prevalence or treatment. |
| Entity Errors          | Wrong factual details like numbers, names, or dosages.                |
| Contradiction          | Summary directly opposes the abstract's results or claims.            |
| Omission               | Missing key findings or results from the abstract.                    |
| Jumping to Conclusions | Overstating results beyond what data supports.                        |
| Misinterpretation      | Misstating or oversimplifying meaning of the abstract.                |
| Structure Error        | Disorganized layout or mixing sections like methods and results.      |
| Persona Relevance      | Language complexity unsuitable for target persona.                    |
| Hallucination          | Adding fabricated or irrelevant content not in abstract.              |

Table 4: Error types and descriptions for persona-aware biomedical summarization.

1. Each evaluator independently reviews the abstract, the reference lay summary, and the model-generated summaries.
2. Evaluators rate each summary using the specified evaluation facets and the provided Likert scale.

| Method Category          | Method                | Model   | R-1    | R-2    | R-L    | SARI    |
|--------------------------|-----------------------|---------|--------|--------|--------|---------|
| Single LLM               | Zero-shot             | GPT-4o  | 0.5916 | 0.2380 | 0.3434 | 49.5828 |
|                          |                       | Gemini  | 0.5960 | 0.2495 | 0.3424 | 48.3962 |
|                          |                       | Mixtral | 0.5819 | 0.2629 | 0.3622 | 49.3190 |
|                          |                       | Llama3  | 0.5900 | 0.2568 | 0.3429 | 49.0921 |
|                          | Few-shot              | GPT-4o  | 0.6089 | 0.2550 | 0.3629 | 50.4423 |
|                          |                       | Gemini  | 0.6024 | 0.2683 | 0.3648 | 48.3962 |
|                          |                       | Mixtral | 0.5911 | 0.2686 | 0.3688 | 49.2427 |
|                          |                       | Llama3  | 0.6149 | 0.2926 | 0.3823 | 51.9053 |
|                          | Self-Refine           | GPT-4o  | 0.5624 | 0.2100 | 0.3126 | 47.8304 |
|                          |                       | Gemini  | 0.5564 | 0.2084 | 0.3011 | 46.4857 |
|                          |                       | Mixtral | 0.4980 | 0.2028 | 0.2800 | 46.6206 |
|                          |                       | Llama3  | 0.4973 | 0.2073 | 0.2748 | 48.7969 |
| Single LLM + Multi-Agent | Debate + Judge        | GPT-4o  | 0.5121 | 0.1724 | 0.2626 | 45.8277 |
|                          |                       | Gemini  | 0.5135 | 0.1940 | 0.2740 | 47.0927 |
|                          |                       | Mixtral | 0.5031 | 0.1999 | 0.2742 | 45.3938 |
|                          |                       | Llama3  | 0.5081 | 0.1991 | 0.2706 | 46.6280 |
| Multi LLM + Multi-Agent  | Feedback + Judge      |         | 0.4422 | 0.1610 | 0.2354 | 42.5489 |
|                          | Feedback + Voting     |         | 0.5499 | 0.2064 | 0.2918 | 46.7492 |
|                          | Feedback + Combine    |         | 0.5962 | 0.2657 | 0.3469 | 50.7323 |
|                          | Wiki-Feedback + Judge |         | 0.5691 | 0.2344 | 0.3189 | 45.8352 |
|                          | Error Type + Judge    |         | 0.5730 | 0.2317 | 0.3236 | 46.2927 |
|                          | Error Type + Voting   |         | 0.5680 | 0.2284 | 0.3204 | 45.9240 |
|                          | Error Type + Combine  |         | 0.5756 | 0.2352 | 0.3263 | 46.7815 |
|                          | Error Based + Judge   |         | 0.5653 | 0.2370 | 0.3143 | 45.3778 |
|                          | Error Based + Voting  |         | 0.5601 | 0.2331 | 0.3108 | 44.9326 |
|                          | Error Based + Combine |         | 0.5669 | 0.2396 | 0.3161 | 45.8129 |

Table 5: Comprehensiveness scores for Pre-med Persona



| Method Category          | Method                | Model   | FKGL  | DCRS  | CLI   | LENS    |
|--------------------------|-----------------------|---------|-------|-------|-------|---------|
| Single LLM               | Zero-shot             | GPT-4o  | 17.91 | 12.77 | 19.44 | 46.2652 |
|                          |                       | Gemini  | 15.47 | 12.69 | 18.34 | 49.8001 |
|                          |                       | Mixtral | 16.08 | 11.80 | 17.53 | 50.3914 |
|                          |                       | Llama3  | 17.32 | 12.28 | 18.21 | 49.5902 |
|                          | Few-shot              | GPT-4o  | 15.74 | 11.12 | 16.73 | 45.8277 |
|                          |                       | Gemini  | 15.43 | 12.57 | 18.03 | 53.2864 |
|                          |                       | Mixtral | 16.04 | 11.80 | 17.46 | 48.1884 |
|                          |                       | Llama3  | 16.72 | 11.78 | 17.69 | 53.6619 |
|                          | Self-Refine           | GPT-4o  | 15.83 | 11.43 | 17.59 | 47.0772 |
|                          |                       | Gemini  | 14.82 | 10.76 | 16.42 | 45.4320 |
|                          |                       | Mixtral | 15.42 | 10.52 | 17.01 | 45.9425 |
|                          |                       | Llama3  | 16.42 | 10.07 | 15.98 | 46.0638 |
| Single LLM + Multi-Agent | Debate + Judge        | GPT-4o  | 13.85 | 10.70 | 15.83 | 49.7191 |
|                          |                       | Gemini  | 14.57 | 11.01 | 16.26 | 49.5062 |
|                          |                       | Mixtral | 14.15 | 10.23 | 15.72 | 48.0412 |
|                          |                       | Llama3  | 14.82 | 10.22 | 15.62 | 47.7128 |
| Multi LLM + Multi-Agent  | Feedback + Judge      |         | 12.67 | 10.19 | 14.26 | 48.9614 |
|                          | Feedback + Voting     |         | 14.51 | 10.50 | 15.60 | 48.8772 |
|                          | Feedback + Combine    |         | 14.96 | 10.35 | 15.72 | 50.1168 |
|                          | Wiki-Feedback + Judge |         | 15.15 | 10.58 | 16.24 | 47.3505 |
|                          | Error Type + Judge    |         | 16.14 | 10.52 | 16.62 | 49.5623 |
|                          | Error Type + Voting   |         | 15.80 | 10.40 | 16.40 | 49.0148 |
|                          | Error Type + Combine  |         | 16.25 | 10.60 | 16.75 | 50.0431 |
|                          | Error Based + Judge   |         | 14.15 | 10.63 | 15.85 | 45.1387 |
|                          | Error Based + Voting  |         | 13.95 | 10.55 | 15.65 | 44.7042 |
|                          | Error Based + Combine |         | 14.30 | 10.70 | 15.95 | 45.6129 |

Table 6: Readability Scores for Pre-med Persona

| Method Category          | Method                | Model   | R-1    | R-2    | R-L    | SARI    |
|--------------------------|-----------------------|---------|--------|--------|--------|---------|
| Single LLM               | Zero-shot             | GPT-4o  | 0.6207 | 0.2742 | 0.3831 | 43.9918 |
|                          |                       | Gemini  | 0.6093 | 0.2838 | 0.3881 | 43.1912 |
|                          |                       | Mixtral | 0.6165 | 0.3039 | 0.4086 | 49.5102 |
|                          |                       | Llama3  | 0.6120 | 0.2880 | 0.3719 | 48.4270 |
|                          | Few-shot              | GPT-4o  | 0.6259 | 0.2778 | 0.3878 | 49.1682 |
|                          |                       | Gemini  | 0.6242 | 0.2956 | 0.4000 | 44.9244 |
|                          |                       | Mixtral | 0.6122 | 0.3072 | 0.3978 | 49.4988 |
|                          |                       | Llama3  | 0.6164 | 0.2987 | 0.3838 | 48.2722 |
|                          | Self-Refine           | GPT-4o  | 0.5624 | 0.2100 | 0.3126 | 46.3540 |
|                          |                       | Gemini  | 0.5564 | 0.2084 | 0.3011 | 46.6575 |
|                          |                       | Mixtral | 0.4980 | 0.2028 | 0.2800 | 46.6433 |
|                          |                       | Llama3  | 0.4973 | 0.2073 | 0.2748 | 48.1174 |
| Single LLM + Multi-Agent | Debate + Judge        | GPT-4o  | 0.5521 | 0.1982 | 0.2905 | 43.9918 |
|                          |                       | Gemini  | 0.5792 | 0.2382 | 0.3308 | 43.1912 |
|                          |                       | Mixtral | 0.5606 | 0.2392 | 0.3166 | 44.1105 |
|                          |                       | Llama3  | 0.5380 | 0.2192 | 0.2827 | 46.2658 |
| Multi LLM + Multi-Agent  | Feedback + Judge      |         | 0.5570 | 0.2161 | 0.3006 | 45.7067 |
|                          | Feedback + Voting     |         | 0.5610 | 0.2182 | 0.3045 | 45.8344 |
|                          | Feedback + Combine    |         | 0.6106 | 0.2815 | 0.3677 | 49.8121 |
|                          | Wiki-Feedback + Judge |         | 0.5651 | 0.2283 | 0.3128 | 45.7165 |
|                          | Error Type + Judge    |         | 0.5730 | 0.2317 | 0.3236 | 45.8457 |
|                          | Error Type + Voting   |         | 0.5680 | 0.2284 | 0.3204 | 45.4820 |
|                          | Error Type + Combine  |         | 0.5756 | 0.2352 | 0.3263 | 46.3290 |
|                          | Error Based + Judge   |         | 0.5653 | 0.2370 | 0.3143 | 45.5971 |
|                          | Error Based + Voting  |         | 0.5601 | 0.2331 | 0.3108 | 45.1187 |
|                          | Error Based + Combine |         | 0.5669 | 0.2396 | 0.3161 | 45.9640 |

Table 7: Comprehensiveness scores for Re-searcher Persona

| Method Category          | Method                | Model   | FKGL  | DCRS  | CLI   | LENS    |
|--------------------------|-----------------------|---------|-------|-------|-------|---------|
| Single LLM               | Zero-shot             | GPT-4o  | 15.87 | 11.00 | 16.48 | 51.2367 |
|                          |                       | Gemini  | 15.53 | 11.19 | 16.88 | 53.5571 |
|                          |                       | Mixtral | 15.34 | 10.79 | 16.39 | 55.6439 |
|                          |                       | Llama3  | 15.19 | 10.02 | 14.95 | 57.7103 |
|                          | Few-shot              | GPT-4o  | 16.14 | 11.10 | 16.86 | 59.1560 |
|                          |                       | Gemini  | 15.28 | 11.01 | 16.42 | 54.1410 |
|                          |                       | Mixtral | 15.05 | 10.65 | 15.93 | 59.3144 |
|                          |                       | Llama3  | 15.40 | 10.14 | 15.23 | 60.6989 |
|                          | Self-Refine           | GPT-4o  | 15.83 | 11.43 | 17.59 | 51.3194 |
|                          |                       | Gemini  | 14.82 | 10.76 | 16.42 | 50.6804 |
|                          |                       | Mixtral | 15.42 | 10.52 | 17.01 | 51.2360 |
|                          |                       | Llama3  | 16.42 | 10.07 | 15.98 | 49.0160 |
| Single LLM + Multi-Agent | Debate + Judge        | GPT-4o  | 15.81 | 11.27 | 17.55 | 52.7774 |
|                          |                       | Gemini  | 14.93 | 11.03 | 16.53 | 48.8293 |
|                          |                       | Mixtral | 15.80 | 10.63 | 17.21 | 50.2164 |
|                          |                       | Llama3  | 17.33 | 10.34 | 16.55 | 48.5189 |
| Multi LLM + Multi-Agent  | Feedback + Judge      |         | 15.79 | 10.80 | 16.95 | 52.7070 |
|                          | Feedback + Voting     |         | 15.74 | 10.92 | 16.92 | 52.0788 |
|                          | Feedback + Combine    |         | 16.21 | 11.01 | 16.99 | 52.0884 |
|                          | Wiki-Feedback + Judge |         | 16.08 | 10.99 | 17.28 | 53.6723 |
|                          | Error Type + Judge    |         | 15.87 | 11.19 | 17.49 | 55.5326 |
|                          | Error Type + Voting   |         | 15.78 | 11.08 | 17.40 | 55.0210 |
|                          | Error Type + Combine  |         | 15.95 | 11.22 | 17.55 | 55.8870 |
|                          | Error Based + Judge   |         | 16.58 | 10.98 | 17.40 | 52.7774 |
|                          | Error Based + Voting  |         | 16.40 | 10.92 | 17.32 | 52.3012 |
|                          | Error Based + Combine |         | 16.62 | 11.04 | 17.46 | 53.0585 |

Table 8: Readability scores for Researcher Persona

| Method Category          | Method                | Model   | R-1    | R-2    | R-L    | SARI    |
|--------------------------|-----------------------|---------|--------|--------|--------|---------|
| Single LLM               | Zero-shot             | GPT-4o  | 0.5972 | 0.2439 | 0.3390 | 53.0256 |
|                          |                       | Gemini  | 0.6029 | 0.2624 | 0.3691 | 53.0180 |
|                          |                       | Mixtral | 0.5235 | 0.2068 | 0.3186 | 48.3885 |
|                          |                       | Llama3  | 0.5900 | 0.2585 | 0.3527 | 52.0627 |
|                          | Few-shot              | GPT-4o  | 0.6016 | 0.2476 | 0.3554 | 53.5134 |
|                          |                       | Gemini  | 0.6018 | 0.2611 | 0.3773 | 53.2864 |
|                          |                       | Mixtral | 0.5073 | 0.1892 | 0.3014 | 47.2191 |
|                          |                       | Llama3  | 0.5841 | 0.2542 | 0.3499 | 53.6619 |
|                          | Self-Refine           | GPT-4o  | 0.5288 | 0.1727 | 0.2696 | 48.3749 |
|                          |                       | Gemini  | 0.5422 | 0.1874 | 0.2756 | 47.9891 |
|                          |                       | Mixtral | 0.4750 | 0.1609 | 0.2523 | 47.1213 |
|                          |                       | Llama3  | 0.4233 | 0.1366 | 0.2008 | 47.4832 |
| Single LLM + Multi-Agent | Debate + Judge        | GPT-4o  | 0.4614 | 0.1212 | 0.2217 | 46.4322 |
|                          |                       | Gemini  | 0.2751 | 0.0268 | 0.1274 | 37.2884 |
|                          |                       | Mixtral | 0.4341 | 0.1316 | 0.2237 | 48.3885 |
|                          |                       | Llama3  | 0.2724 | 0.0288 | 0.1269 | 37.7294 |
| Multi LLM + Multi-Agent  | Feedback + Judge      |         | 0.4721 | 0.1380 | 0.2344 | 46.6069 |
|                          | Feedback + Voting     |         | 0.4810 | 0.1489 | 0.2402 | 47.0179 |
|                          | Feedback + Combine    |         | 0.5478 | 0.2067 | 0.2914 | 50.6332 |
|                          | Wiki-Feedback + Judge |         | 0.5158 | 0.1729 | 0.2665 | 48.2856 |
|                          | Error Type + Judge    |         | 0.4687 | 0.1350 | 0.2355 | 45.1380 |
|                          | Error Type + Voting   |         | 0.4640 | 0.1322 | 0.2320 | 44.7021 |
|                          | Error Type + Combine  |         | 0.4706 | 0.1384 | 0.2371 | 45.6214 |
|                          | Error Based + Judge   |         | 0.4837 | 0.1390 | 0.2415 | 45.9380 |
|                          | Error Based + Voting  |         | 0.4791 | 0.1361 | 0.2380 | 45.4027 |
|                          | Error Based + Combine |         | 0.4854 | 0.1415 | 0.2430 | 46.3886 |

Table 9: Comprehensiveness scores for Lay Persona

| Method Category          | Method                | Model   | FKGL  | DCRS  | CLI   | LENS    |
|--------------------------|-----------------------|---------|-------|-------|-------|---------|
| Single LLM               | Zero-shot             | GPT-4o  | 10.37 | 8.54  | 11.30 | 51.7573 |
|                          |                       | Gemini  | 10.04 | 8.71  | 11.66 | 52.5033 |
|                          |                       | Mixtral | 12.38 | 9.37  | 12.93 | 62.7845 |
|                          |                       | Llama3  | 12.07 | 8.59  | 11.86 | 53.1642 |
|                          | Few-shot              | GPT-4o  | 11.22 | 8.94  | 11.99 | 75.0276 |
|                          |                       | Gemini  | 11.00 | 8.97  | 12.17 | 69.1026 |
|                          |                       | Mixtral | 13.08 | 9.81  | 13.84 | 64.7531 |
|                          |                       | Llama3  | 12.06 | 8.70  | 11.99 | 78.4790 |
|                          | Self-Refine           | GPT-4o  | 11.24 | 9.47  | 12.84 | 71.6981 |
|                          |                       | Gemini  | 9.44  | 8.45  | 11.14 | 71.5003 |
|                          |                       | Mixtral | 12.49 | 9.55  | 13.71 | 57.7433 |
|                          |                       | Llama3  | 11.84 | 8.32  | 12.07 | 62.4744 |
| Single LLM + Multi-Agent | Debate + Judge        | GPT-4o  | 15.15 | 13.08 | 16.81 | 51.7573 |
|                          |                       | Gemini  | 12.54 | 11.86 | 14.15 | 52.5033 |
|                          |                       | Mixtral | 14.70 | 12.72 | 15.57 | 57.8722 |
|                          |                       | Llama3  | 15.55 | 12.09 | 14.80 | 53.1642 |
| Multi LLM + Multi-Agent  | Feedback + Judge      |         | 14.64 | 12.73 | 15.87 | 54.6001 |
|                          | Feedback + Voting     |         | 14.56 | 12.50 | 15.35 | 58.2415 |
|                          | Feedback + Combine    |         | 12.80 | 11.54 | 13.60 | 75.6658 |
|                          | Wiki-Feedback + Judge |         | 13.66 | 12.19 | 14.93 | 66.8700 |
|                          | Error Type + Judge    |         | 14.32 | 12.63 | 15.85 | 56.5742 |
|                          | Error Type + Voting   |         | 14.18 | 12.42 | 15.62 | 56.0216 |
|                          | Error Type + Combine  |         | 14.40 | 12.71 | 15.94 | 56.9820 |
|                          | Error Based + Judge   |         | 14.52 | 12.83 | 16.25 | 58.0742 |
|                          | Error Based + Voting  |         | 14.39 | 12.62 | 16.02 | 57.4185 |
|                          | Error Based + Combine |         | 14.61 | 12.90 | 16.31 | 58.6034 |

Table 10: Readability scores for Lay Persona

| Method Category          | Method                | Model   | R-1    | R-2    | R-L    | SARI    |
|--------------------------|-----------------------|---------|--------|--------|--------|---------|
| Single LLM               | Zero-shot             | GPT-4o  | 0.6000 | 0.2905 | 0.3997 | 42.1944 |
|                          |                       | Gemini  | 0.6348 | 0.3451 | 0.4511 | 45.1591 |
|                          |                       | Mixtral | 0.6478 | 0.3911 | 0.4758 | 48.2700 |
|                          |                       | Llama3  | 0.6594 | 0.3928 | 0.4738 | 47.8625 |
|                          | Few-shot              | GPT-4o  | 0.5418 | 0.2416 | 0.3555 | 41.4523 |
|                          |                       | Gemini  | 0.6599 | 0.3815 | 0.4912 | 44.6463 |
|                          |                       | Mixtral | 0.6625 | 0.4035 | 0.4946 | 48.4338 |
|                          |                       | Llama3  | 0.6769 | 0.4198 | 0.5037 | 45.4880 |
|                          | Self-Refine           | GPT-4o  | 0.5577 | 0.2547 | 0.3535 | 42.4489 |
|                          |                       | Gemini  | 0.6010 | 0.3126 | 0.4184 | 44.1335 |
|                          |                       | Mixtral | 0.6552 | 0.3973 | 0.4852 | 48.3519 |
|                          |                       | Llama3  | 0.4732 | 0.2345 | 0.2981 | 43.1135 |
| Single LLM + Multi-Agent | Debate + Judge        | GPT-4o  | 0.5418 | 0.2416 | 0.3555 | 42.1944 |
|                          |                       | Gemini  | 0.6599 | 0.3815 | 0.4912 | 41.1892 |
|                          |                       | Mixtral | 0.6625 | 0.4035 | 0.4946 | 43.3271 |
|                          |                       | Llama3  | 0.6769 | 0.4198 | 0.5037 | 30.5329 |
| Multi LLM + Multi-Agent  | Feedback + Judge      |         | 0.5954 | 0.2974 | 0.3924 | 43.0934 |
|                          | Feedback + Voting     |         | 0.5946 | 0.3007 | 0.3920 | 43.5686 |
|                          | Feedback + Combine    |         | 0.6148 | 0.3576 | 0.4213 | 47.8559 |
|                          | Wiki-Feedback + Judge |         | 0.6057 | 0.3048 | 0.3975 | 42.3177 |
|                          | Error Type + Judge    |         | 0.6012 | 0.2884 | 0.3938 | 40.5707 |
|                          | Error Type + Voting   |         | 0.5964 | 0.2852 | 0.3905 | 40.1189 |
|                          | Error Type + Combine  |         | 0.6028 | 0.2920 | 0.3953 | 41.0024 |
|                          | Error Based + Judge   |         | 0.5827 | 0.2843 | 0.3633 | 41.7574 |
|                          | Error Based + Voting  |         | 0.5782 | 0.2812 | 0.3601 | 41.2036 |
|                          | Error Based + Combine |         | 0.5841 | 0.2865 | 0.3650 | 42.1298 |

Table 11: Comprehensiveness scores for Expert Persona

| Method Category          | Method                | Model   | FKGL  | DCRS  | CLI   | LENS    |
|--------------------------|-----------------------|---------|-------|-------|-------|---------|
| Single LLM               | Zero-shot             | GPT-4o  | 17.91 | 12.77 | 19.44 | 51.2367 |
|                          |                       | Gemini  | 15.47 | 12.69 | 18.34 | 53.5571 |
|                          |                       | Mixtral | 16.08 | 11.80 | 17.53 | 55.6439 |
|                          |                       | Llama3  | 17.32 | 12.28 | 18.21 | 57.7103 |
|                          | Few-shot              | GPT-4o  | 15.57 | 11.39 | 16.84 | 59.1560 |
|                          |                       | Gemini  | 15.43 | 12.57 | 18.03 | 54.1410 |
|                          |                       | Mixtral | 16.04 | 11.80 | 17.46 | 59.3144 |
|                          |                       | Llama3  | 16.72 | 11.78 | 17.69 | 60.6989 |
|                          | Self-Refine           | GPT-4o  | 17.30 | 12.76 | 19.39 | 51.3194 |
|                          |                       | Gemini  | 15.36 | 12.49 | 17.91 | 50.6804 |
|                          |                       | Mixtral | 16.06 | 11.80 | 17.50 | 57.4792 |
|                          |                       | Llama3  | 17.33 | 11.95 | 18.10 | 49.0160 |
| Single LLM + Multi-Agent | Debate + Judge        | GPT-4o  | 15.57 | 11.39 | 16.84 | 53.5064 |
|                          |                       | Gemini  | 15.43 | 12.57 | 18.03 | 48.8293 |
|                          |                       | Mixtral | 16.04 | 11.80 | 17.46 | 50.2164 |
|                          |                       | Llama3  | 16.72 | 11.78 | 17.69 | 48.5189 |
| Multi LLM + Multi-Agent  | Feedback + Judge      |         | 15.44 | 12.18 | 17.78 | 52.7070 |
|                          | Feedback + Voting     |         | 16.34 | 12.39 | 18.49 | 52.0788 |
|                          | Feedback + Combine    |         | 17.24 | 11.98 | 18.26 | 52.0884 |
|                          | Wiki-Feedback + Judge |         | 16.45 | 12.31 | 18.44 | 52.8652 |
|                          | Error Type + Judge    |         | 17.23 | 12.57 | 19.10 | 55.5326 |
|                          | Error Type + Voting   |         | 17.10 | 12.42 | 18.92 | 55.0214 |
|                          | Error Type + Combine  |         | 17.28 | 12.60 | 19.14 | 55.8871 |
|                          | Error Based + Judge   |         | 17.10 | 12.37 | 18.90 | 52.7774 |
|                          | Error Based + Voting  |         | 16.97 | 12.22 | 18.72 | 52.3189 |
|                          | Error Based + Combine |         | 17.14 | 12.40 | 18.96 | 53.0642 |
|                          |                       |         |       |       |       |         |
|                          |                       |         |       |       |       |         |

Table 12: Readability scores for Expert Persona

| Method      | Model    | Tokens | Cost (USD) |
|-------------|----------|--------|------------|
| Few-shot    | LLaMA-3  | 3.5K   | \$0.0010   |
|             | Mixtral  | 4.2K   | \$0.0020   |
|             | Gemini   | 3.5K   | \$0.0003   |
|             | GPT-4o   | 3.2K   | \$0.0104   |
| Self-Refine | LLaMA-3  | 6.3K   | \$0.0020   |
|             | Mixtral  | 8.7K   | \$0.0049   |
|             | Gemini   | 5.5K   | \$0.0007   |
|             | GPT-4o   | 6.7K   | \$0.0218   |
| MAPS        | 4 agents | 49.7K  | \$0.0575   |

Table 13: Average token usage and estimated cost per sample across methods and models, estimated on 10 samples per persona (40 total).

## B. Researcher Persona Prompts

### Stage 1: Initial Generation

You are a knowledgeable science communicator with interdisciplinary expertise. Your task is to write a clear and comprehensive summary of a scientific abstract for researchers outside the specific field of the study but with a solid understanding of general scientific principles.

When summarizing, follow these key principles:

The audience consists of researchers in fields other than biology and medicine who may not be familiar with domain-specific terminology or methods. Use clear and accessible language. Avoid unnecessary technical jargon; when technical terms are necessary, explain them briefly in context. Accurately represent the study's main findings, methodology, and results. Maintain a professional but approachable tone. Aim for clarity and precision without overly simple or domain-specific language. Structure the summary as a paragraph. Do not use bullet points, numbered lists, or Q&A formats. Focus on the scientific content only. Do not include commentary about the summarization process or subjective judgment of the study's importance.

The final summary should not exceed 350 words.

#### ABSTRACT

{abstract}

### Stage 2: Persona-specific Feedback

**Task:** Review the provided summary compared to the original abstract. Suggest improvements to make the summary better intended for researchers with a general scientific background who are not specialists in the specific field.

Feedback categories include definition enhancement, synonym suggestions, simplification needs, detail adjustment, factuality check, overall coherence, and key point highlighting.

For each category, provide specific examples from the summary and propose clear, constructive improvements. Where applicable, explain why the changes would enhance understanding for non-specialist researchers.

#### ABSTRACT

{abstract}

#### SUMMARY

{summary}

### Stage 3: Feedback-Guided Revision

Refine the following summary intended for researchers with a general scientific background, but who are not specialists in the specific field, based on the feedback provided below.

Do not use numbered lists to explain terminology, methods, or findings, as this negatively impacts readability. Ensure that the revised summary does not exceed 350 words.

#### SUMMARY

{last\_summary}

#### FEEDBACK

{feedback}

## B.1. Adjudication Prompts for the Researcher Persona

### Judge Strategy

You are a knowledgeable science communicator with strong interdisciplinary expertise evaluating summaries of a scientific abstract. Each summary is intended for researchers outside the original field but with a solid scientific background.

Evaluate each summary based on accuracy and faithfulness, clarity and accessibility, comprehensiveness, and professional tone. After reviewing each summary step by step, report your reasoning and state your final judgment as **Summary X**.

#### ABSTRACT

{abstract}

#### SUMMARIES

{summaries\_text}

### Combine Strategy

You are a skilled science communicator with deep interdisciplinary expertise. Your task is to produce a single clear, accurate, and accessible summary for non-medical researchers using multiple candidate summaries as input.

The final summary must maintain scientific faithfulness, avoid excessive jargon, and clearly explain findings and methods to non-specialists.

#### ABSTRACT

{abstract}

#### CANDIDATE SUMMARIES

{summaries\_text}



# Automated Detection of Dosing Errors in Clinical Trial Narratives: A Multi-Modal Feature Engineering Approach with LightGBM

Mohammad AL-Smadi

Qatar University  
Doha, Qatar  
malsmadi@qu.edu.qa

## Abstract

Clinical trials require strict adherence to medication protocols, yet dosing errors remain a persistent challenge affecting patient safety and trial integrity. We present an automated system for detecting dosing errors in unstructured clinical trial narratives using gradient boosting with comprehensive multi-modal feature engineering. Our approach combines 3,451 features spanning traditional NLP (TF-IDF, character n-grams), dense semantic embeddings (all-MiniLM-L6-v2), domain-specific medical patterns, and transformer-based scores (BiomedBERT, DeBERTa-v3), used to train a LightGBM model. Features are extracted from nine complementary text fields (median 5,400 characters per sample) ensuring complete coverage across all 42,112 clinical trial narratives. On the CT-DEB benchmark dataset with severe class imbalance (4.9% positive rate), we achieve 0.8725 test ROC-AUC through 5-fold ensemble averaging (cross-validation:  $0.8833 \pm 0.0091$  AUC). Systematic ablation studies reveal that removing sentence embeddings causes the largest performance degradation (2.39%), demonstrating their critical role despite contributing only 37.07% of total feature importance. Feature efficiency analysis demonstrates that selecting the top 500-1000 features yields optimal performance (0.886-0.887 AUC), outperforming the full 3,451-feature set (0.879 AUC) through effective noise reduction. Our findings highlight the importance of feature selection as a regularization technique and demonstrate that sparse lexical features remain complementary to dense representations for specialized clinical text classification under severe class imbalance.

**Keywords:** Clinical Trial Narratives, Dosing Error Detection, Patient Safety, Natural Language Processing, Multi-Modal Feature Engineering, LightGBM, Transformer Models, Medical Text Classification

## 1. Introduction

Clinical trials are fundamental to pharmaceutical development and medical advancement, requiring strict adherence to pre-defined protocols specifying medication dosing, timing, and administration routes. Dosing errors—deviations from these protocols—pose significant risks to patient safety and trial validity (ICH Expert Working Group, 2001; Nakayama, 2007; Cingi and Muluk, 2017). Such errors can result in adverse events, compromise study endpoints, and lead to regulatory issues or trial termination.

Manual review of clinical trial documentation is the current standard for identifying protocol deviations. However, with modern trials generating thousands of narratives describing patient visits and medication administration, this approach is time-consuming, expensive, and subject to human error. For instance, a typical Phase III trial may involve 1,000+ patients across multiple sites, each generating dozens of clinical narratives throughout the study period. Manual review of this volume overwork quality assurance teams and may miss subtle deviations. (Cingi and Muluk, 2017)

Clinical Trials Dosing Errors Benchmark 2026 (CT-DEB'26) (Ferdowsi et al., 2026) aims at addressing these challenges in the review of clinical trial documentation with machine learning. Natural language processing (NLP) offers potential for

automating dosing error detection. However, the task presents unique challenges: (1) Severe class imbalance—only 4-5% of narratives contain errors; (2) Complex medical language—specialized terminology, abbreviations, and implicit references; (3) Subtle linguistic cues—protocol deviations often expressed indirectly through phrases like “dose adjusted per investigator discretion”; (4) Variable text structure—narratives range from 50 to 5,000+ characters with inconsistent formatting; (5) Context dependency—distinguishing planned dose modifications from unplanned deviations requires understanding protocol context.

This paper presents an automated dosing error detection system using multi-modal feature engineering and gradient boosting. Our key contributions are:

1. A 3,451-dimensional feature space combining traditional NLP (Term Frequency–Inverse Document Frequency (TF-IDF), character n-grams), sentence embeddings, medical patterns, and transformer-based scores.
2. Systematic analysis showing sentence embeddings and lexical features are complementary, while transformer scores underperform.
3. Optuna-optimized 5-fold ensemble achieving  $0.8833 \pm 0.0091$  CV AUC with minimal overfitting (0.69% out-of-fold(OOF)-test gap).

4. Feature selection improves performance: 500-1000 features (14-29%) achieve 0.886-0.887 AUC, outperforming the full baseline (0.879) through noise reduction.
5. Test ROC-AUC of 0.8725 on CT-DEB with threshold-adjustable recall (26-60%) for flexible deployment.

## 2. Related Work

### 2.1. Clinical NLP and Information Extraction

Clinical NLP has evolved significantly over the past two decades. Early systems like cTAKES (Savova et al., 2010) and MetaMap (Aronson and Lang, 2010) provided rule-based approaches for medical concept extraction and normalization. The i2b2 shared tasks (Uzuner et al., 2011) established benchmarks for concept extraction, assertion detection, and relation extraction from clinical texts, demonstrating that machine learning approaches could achieve substantial performance gains.

More recent work has focused on applying deep learning to clinical texts. Jagannatha and Yu (2016) demonstrated that recurrent neural networks with structured prediction models improve sequence labeling in clinical narratives. Si et al. (2019) showed that contextualized embeddings enhance clinical concept extraction, achieving state-of-the-art results on medical entity recognition tasks.

### 2.2. Biomedical Language Models

The introduction of BERT (Devlin et al., 2019) revolutionized NLP, and several domain-specific variants have been developed for biomedicine. BioBERT (Lee et al., 2020) continues pre-training BERT on biomedical corpora (PubMed abstracts and PMC articles), achieving improvements on various biomedical NLP tasks. Gu et al. (2021) challenged the assumption that domain-specific models benefit from starting with general-domain weights, showing that pre-training from scratch on biomedical text (PubMedBERT/BiomedBERT) yields superior performance. Clinical-specific BERT variants (Alsentzer et al., 2019) trained on clinical notes from MIMIC-III (Johnson et al., 2016) have shown promise for clinical tasks.

### 2.3. Medical Error Detection

Adverse drug event (ADE) detection has received considerable attention. Harpaz et al. (2012) developed data-mining methodologies for ADE discovery. These approaches typically focus on detecting harmful outcomes rather than protocol compliance.

Clinical trial eligibility screening has been explored by Kalankesh and Monaghesh (2024), who used electronic health records to identify patients meeting trial criteria. While related, eligibility screening differs fundamentally from protocol deviation detection—the former matches patients to protocols, while the latter identifies deviations from assigned protocols.

Ferdowsi et al. (2023) applied deep learning to predict clinical trial outcomes based on protocol design features, achieving strong performance in identifying trials at risk of failure. Their work demonstrates the value of automated analysis of clinical trial documentation, though it focuses on trial-level risk prediction rather than individual dosing error detection in narrative text.

### 2.4. Research Gap

Despite progress in clinical NLP and medical error detection, automated protocol deviation detection in clinical trials remains understudied. Prior work on clinical error detection has primarily leveraged structured EHR data (Rajkomar et al., 2018). Churpek et al. (2016) utilized vital sign trends to predict clinical deterioration on hospital wards, while Zimolzak et al. (2024) applied machine learning to structured variables including demographics, laboratory values, vital signs, orders, and visit times for diagnostic error detection. Similarly, clinical trial data validation typically focuses on structured fields such as laboratory values and vital signs (Yuan et al., 2024).

Substantial NLP research has focused on adverse drug event (ADE) detection from clinical narratives. Li et al. (2018) developed deep learning models for extracting ADEs from EHR notes, while Jagannatha et al. (2019) organized the MADE challenge for medication and ADE extraction. These approaches employ named entity recognition and relation extraction to identify drug-ADE relationships, but focus on detecting harmful reactions rather than protocol compliance in clinical trials.

Limited work exists on protocol deviations in clinical trials. Richard and Reddy (2021) applied TF-IDF and SVM to categorize existing protocol deviation descriptions, enabling trend analysis across trials. However, this work classifies already-identified deviations rather than detecting deviations from unstructured medication administration narratives—a critical distinction for automated quality assurance.

Our work addresses this gap by: (1) Targeting unstructured narrative text describing medication administration events in clinical trials; (2) Handling severe class imbalance typical of quality assurance scenarios (95:5 ratio); (3) Providing systematic feature ablation to understand what drives performance; (4) Demonstrating production-feasible

efficiency through feature selection that improves both accuracy and deployment cost.

### 3. Dataset and Task

#### 3.1. CT-DEB Benchmark

We utilize the CT-DEB (Clinical Trial Dosing Error Benchmark) dataset (Hêche et al., 2026), specifically designed for evaluating automated dosing error detection systems. The dataset comprises clinical trial narratives describing medication administration across various therapeutic areas, protocols, and clinical trial phases.

##### 3.1.1. Dataset Composition

Table 1 summarizes dataset statistics. The dataset comprises 42,112 narratives collected from clinical trial documentation spanning multiple pharmaceutical companies and clinical research organizations (35,794 for training and validation, with 6,318 held out for final testing). Each narrative describes one or more medication administration events for a single patient visit.

| Split        | Total         | Negative       | Positive     |
|--------------|---------------|----------------|--------------|
| Training     | 29,478        | 28,126 (95.4%) | 1,352 (4.6%) |
| Validation   | 6,316         | 6,031 (95.5%)  | 285 (4.5%)   |
| Test         | 6,318         | 6,008 (95.1%)  | 310 (4.9%)   |
| <b>Total</b> | <b>42,112</b> | <b>40,165</b>  | <b>1,947</b> |

Table 1: CT-DEB Dataset Statistics

The severe class imbalance (4.6% positive rate) reflects real-world prevalence of protocol deviations in well-monitored clinical trials.

#### 3.2. Imbalanced Classification

CT-DEB dataset involves severe class imbalance (95:5 negative-to-positive ratio), requiring specialized machine learning techniques. Cost-sensitive learning (Elkan, 2001) and synthetic oversampling (SMOTE) (Chawla et al., 2002) represent classical approaches. Modern gradient boosting frameworks like XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) provide built-in mechanisms for handling imbalance through instance weighting and custom loss functions.

#### 3.3. Task Definition

Formally, given clinical text  $x \in \mathcal{X}$  describing medication administration, we learn function  $f : \mathcal{X} \rightarrow \{0, 1\}$  predicting whether a dosing error occurred. Let  $y = 1$  indicate a protocol deviation and  $y = 0$  indicate proper protocol adherence.

| Category            |      | Dims   | Description                                                                                            |
|---------------------|------|--------|--------------------------------------------------------------------------------------------------------|
| Medical Patterns    | Pat- | 43     | Rule-based features for dose changes, adverse events, text statistics                                  |
| Word TF-IDF         |      | ~2,000 | Unigram features with sublinear TF scaling, L2 normalization, max_features=2,000, min_df=2, max_df=0.8 |
| Char N-grams        |      | ~1,000 | Character sequences (n=3–7), top-1,000                                                                 |
| Sentence Embeddings | Em-  | 386    | all-MiniLM-L6-v2 dense semantics                                                                       |
| Transformer Scores  |      | 2      | BiomedBERT & DeBERTa probabilities                                                                     |

Table 2: Feature Category Overview

The primary evaluation metric is ROC-AUC (area under the receiver operating characteristic curve), chosen for its robustness to class imbalance and ability to evaluate performance across classification thresholds. Secondary metrics include F1-macro, precision, recall, and balanced accuracy.

## 4. Methodology

#### 4.1. Data Preparation and Feature Engineering Pipeline

Our feature engineering pipeline transforms raw clinical narratives into a 3,451-dimensional feature vector combining multiple representation types. Table 2 provides an overview of extracted features.

Clinical trial registry data exhibits inherent sparsity across structured fields. To ensure comprehensive text coverage for all samples, we concatenate nine complementary text fields from the CT-DEB dataset: `briefSummary` (100% coverage), `detailedDescription` (62% coverage), `protocolPdfText` (42% coverage), `armDescriptions` (99% coverage), `interventionDescriptions` (100% coverage), `interventionNames` (100% coverage), `conditions` (100% coverage), `conditionsKeywords` (64% coverage), and `locationDetails` (94% coverage).

For each sample, we concatenate all available non-null text from these nine fields, filtering empty strings and null values. This multi-field approach addresses the sparsity of individual fields (e.g., `protocolPdfText` present in only 42% of records) while ensuring every sample has substantial narrative content. The resulting concatenated text has a median length of approximately 5,400 characters per sample, with all samples (100%) containing at least 200 characters and a minimum observed length of 209 characters. No samples

result in zero-valued feature vectors due to complementary field availability patterns.

The concatenated text serves as input to all text-based feature extractors: word TF-IDF vectorization, character n-gram extraction, sentence embedding generation, and transformer-based scoring.

#### 4.1.1. Medical Pattern Features

We extract 43 handcrafted features using regular expressions applied to the concatenated narrative text, organized into ten subcategories. Table 3 presents the complete taxonomy with feature names and descriptions.

#### 4.1.2. Word TF-IDF Features

We compute Term Frequency-Inverse Document Frequency vectors (Salton and Buckley, 1988) to capture discriminative medical vocabulary. Configuration: vocabulary size limited to 2,000 most informative unigrams, minimum document frequency of 2, maximum document frequency of 0.8 (exclude overly common terms), L2 normalization, sublinear term frequency scaling ( $1 + \log(\text{tf})$ ). This yields approximately 2,000 word-based features that capture key medical terminology and dosing-related vocabulary across the corpus.

#### 4.1.3. Character N-gram Features

To capture subword patterns, morphological variations, and medical abbreviations, we extract character-level n-grams with  $n \in \{3, 4, 5, 6, 7\}$ . Maximum features: 1,000 most informative sequences, minimum document frequency: 2.

#### 4.1.4. Sentence Embeddings

We employ the all-MiniLM-L6-v2 sentence transformer (Reimers and Gurevych, 2019), which generates 386-dimensional dense semantic representations of entire narratives. Based on the MiniLM architecture (Wang et al., 2020), this model is trained using contrastive learning on over 1 billion sentence pairs from diverse sources.

The model maps semantically similar sentences to nearby points in embedding space, enabling it to capture: semantic similarity (“adverse event”  $\approx$  “toxicity”  $\approx$  “side effect”), temporal relationships (“before treatment” vs “after treatment”), causal relationships (“due to”, “because of”, “resulting in”), negation (“no dose change” vs “dose change”), certainty/hedging (“definitely” vs “possibly”), and protocol compliance context (“as per protocol” vs “deviation from protocol”).

#### 4.1.5. Transformer-based Probability Scores

We additionally incorporate two domain-specific transformer models fine-tuned on the dosing error detection task:

**BiomedBERT** (Gu et al., 2021): Pre-trained from scratch on 3.17 million PubMed abstracts and 1.04 million full-text articles from PubMed Central, totaling approximately 14GB of biomedical text. We fine-tune the base model (12 layers, 768 hidden dimensions, 110M parameters) on our training data for binary classification.

**DeBERTa-v3-base** (He et al., 2021): Employs disentangled attention mechanism separating content and position representations. The base variant (12 layers, 768 hidden dimensions, 86M parameters) is fine-tuned on our task.

For narratives exceeding 512 tokens (BERT’s maximum sequence length), we employ a sliding window approach: text is split into overlapping 512-token chunks with 128-token overlap, each chunk processed independently. To obtain a trial-level score, we apply *top-k mean pooling* across chunk predictions ( $k=3$ ), averaging the three highest probabilities. This approach emphasizes high-risk segments while mitigating noise from less relevant sections.

The final output from each transformer is a single scalar probability. These two probabilities are appended as structured features to the tabular model. As shown in Section 5, their contribution to overall performance is modest, likely due to information compression when reducing rich contextual embeddings to scalar probabilities.

### 4.2. Model Architecture

We employ LightGBM (Ke et al., 2017), a gradient boosting framework based on decision trees. LightGBM was selected for several reasons: (1) Efficiency with high-dimensional sparse features through histogram-based learning and gradient-based one-side sampling (GOSS); (2) Built-in handling of class imbalance via instance weighting; (3) Interpretability through feature importance analysis; (4) Fast training and inference (minutes on CPU vs hours for deep learning); (5) Strong empirical performance on tabular data.

The model architecture consists of an ensemble of  $T = 4000$  decision trees, where each tree is trained to minimize the residual error of the previous ensemble. The final prediction is computed as:

$$\hat{y} = \sigma \left( \sum_{t=1}^T \eta \cdot f_t(\mathbf{x}) \right) \quad (1)$$

where  $T = 4000$  is the number of trees,  $\eta = 0.01$  is the learning rate (step size shrinkage),  $f_t$  is the  $t$ -th decision tree mapping input features  $\mathbf{x} \in \mathbb{R}^{3451}$



| Category            | # | Features & Description                                                                                                                                                                                                                                                                                   |
|---------------------|---|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Dose Units          | 5 | Binary indicators for dose measurement units: <code>has_mg_dose</code> (milligrams), <code>has_ml_dose</code> (milliliters), <code>has_mcg_dose</code> (micrograms), <code>has_iu_dose</code> (international units), <code>has_unit_dose</code> (generic units)                                          |
| Dose Calculations   | 2 | <code>has_weight_based</code> (mg/kg, mcg/kg dosing), <code>has_bsa_based</code> (body surface area dosing: mg/m <sup>2</sup> )                                                                                                                                                                          |
| Routes              | 6 | Administration route flags: <code>has_iv</code> (intravenous), <code>has_oral</code> (oral), <code>has_sc</code> (subcutaneous), <code>has_im</code> (intramuscular), <code>has_topical</code> (topical), <code>has_inhaled</code> (inhaled)                                                             |
| Frequencies         | 5 | Dosing schedule indicators: <code>has_qd</code> (once daily), <code>has_bid</code> (twice daily), <code>has_tid</code> (three times daily), <code>has_qid</code> (four times daily), <code>has_prn</code> (as needed)                                                                                    |
| Dose Concepts       | 6 | Management patterns: <code>has_max_dose</code> (maximum limits), <code>has_titration</code> (escalation), <code>has_loading_dose</code> , <code>has_maintenance</code> , <code>has_adjustment</code> , <code>has_contraindication</code>                                                                 |
| Special Populations | 5 | Population-specific considerations: <code>has_pediatric</code> , <code>has_geriatric</code> , <code>has_pregnancy</code> , <code>has_renal</code> (renal impairment), <code>has_hepatic</code> (hepatic dysfunction)                                                                                     |
| Error Indicators    | 1 | <code>has_error_keyword</code> : explicit mentions of errors, mistakes, overdoses, underdoses, miscalculations, or deviations                                                                                                                                                                            |
| Count Features      | 4 | Quantitative counts: <code>dose_count</code> , <code>percentage_count</code> , <code>decimal_count</code> , <code>range_count</code>                                                                                                                                                                     |
| Text Statistics     | 4 | Statistical properties: <code>text_length</code> (characters), <code>word_count</code> , <code>sentence_count</code> , <code>avg_word_length</code>                                                                                                                                                      |
| Study Metadata      | 5 | Registry fields: <code>num_trials</code> (typically 1), <code>num_conditions</code> , <code>enrollment_count</code> , <code>phase_encoded</code> (1–4), <code>study_type_encoded</code> (1=interventional, 2=observational). Populated from dataframe columns when available; default to zero otherwise. |

Table 3: Medical Pattern Features Taxonomy (43 features total)

to a real-valued score, and  $\sigma$  is the sigmoid function mapping to probability space  $[0, 1]$ .

### 4.3. Hyperparameter Optimization

We employ Optuna (Akiba et al., 2019), a Bayesian optimization framework, to systematically tune LightGBM hyperparameters. We conducted 50 trials with 5-fold cross-validation per trial, maximizing mean validation ROC-AUC across folds. As presented in Table 4, each trial explored a span of hyperparameter space, the best trial (Trial 18) achieved optimal cross-validation performance with  $0.8833 \pm 0.0091$  ROC-AUC.

Key parameter choices from Optuna optimization:

**Learning rate and iterations:** Low learning rate (0.0054) with many iterations (4,000) allows grad-

ual learning while early stopping (200-iteration patience) prevents overtraining.

**Tree structure:** `num_leaves=118` and `max_depth=9` balance expressiveness with generalization. `min_child_samples=211` ensures leaves represent sufficient examples.

**Regularization:** L1 (4.29) and L2 (4.33) penalties prevent overfitting. Feature/bagging fractions (0.795/0.813) promote ensemble diversity through random subsampling.

**Class balancing:** `scale_pos_weight=20.87` compensates for the 20:1 class imbalance, giving proper weight to rare positive examples during training.



| Parameter                  | Search Space  | Value  | Description                                         |
|----------------------------|---------------|--------|-----------------------------------------------------|
| <i>Learning Parameters</i> |               |        |                                                     |
| n_estimators               | —             | 4,000  | Number of boosting rounds                           |
| learning_rate              | [0.005, 0.05] | 0.0054 | Step size shrinkage, sampled on a logarithmic scale |
| <i>Tree Structure</i>      |               |        |                                                     |
| num_leaves                 | [31, 256]     | 118    | Max leaves per tree                                 |
| max_depth                  | [4, 10]       | 9      | Maximum tree depth                                  |
| min_child_samples          | [20, 300]     | 211    | Min samples per leaf                                |
| <i>Regularization</i>      |               |        |                                                     |
| lambda_l1                  | [0.0, 5.0]    | 4.29   | L1 regularization                                   |
| lambda_l2                  | [0.0, 5.0]    | 4.33   | L2 regularization                                   |
| feature_fraction           | [0.6, 0.9]    | 0.795  | Column sampling ratio                               |
| bagging_fraction           | [0.6, 0.9]    | 0.813  | Row sampling ratio                                  |
| bagging_freq               | —             | 1      | Frequency of subsampling                            |
| <i>Class Imbalance</i>     |               |        |                                                     |
| scale_pos_weight           | —             | 20.87  | Positive class weight                               |

Table 4: LightGBM Hyperparameter Configuration (Optuna Best Trial)

#### 4.4. Training Procedure

Using the Optuna-optimized hyperparameters from previous subsection 4.3, we employ a 5-fold ensemble training strategy<sup>1</sup>:

**Feature Preparation** - All 3,451 features are extracted and stored as scipy sparse matrices in compressed NPZ format (approximately 500MB vs 3GB dense).

**Stratified Cross-Validation** - Train and validation data were combined and then partitioned into 5 stratified folds maintaining class balance (4.6% positive rate per fold). For each fold  $k \in \{1, 2, 3, 4, 5\}$ :

1. Train LightGBM on the remaining 4 folds (80% of data) using weighted binary cross-entropy:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N w_i [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (2)$$

where  $w_i = 20.87$  for positive examples,  $w_i = 1$  for negative.

2. Generate OOF predictions on fold  $k$  (the held-out 20%)
3. Save the trained model for later ensembling

**Out-of-Fold Validation** - Concatenating predictions from all 5 folds yields complete OOF pre-

dictions across the training set, providing an unbiased performance estimate:  $0.8833 \pm 0.0091$  ROC-AUC.

**Ensemble Prediction** - For test inference, all 5 fold models generate predictions, which are averaged:

$$\hat{y}_{\text{ensemble}}(x) = \frac{1}{5} \sum_{k=1}^5 \hat{y}_k(x) \quad (3)$$

## 5. Results

### 5.1. Overall Performance

We employ 5-fold stratified cross-validation with ensemble averaging. Table 5 presents test set performance.

The ROC-AUC of 0.8725 indicates excellent discriminative ability. The ensemble demonstrates stable performance with mean cross-validation AUC of  $0.8833 \pm 0.0091$  across folds, and OOF AUC of 0.8794, showing minimal overfitting (only 0.6% gap to test).

The classification threshold (0.3744) was optimized for F1-score on out-of-fold predictions. This yields 26.1% recall and 33.9% precision—conservative values reflecting the challenge of detecting rare errors in severely imbalanced data (95.1% negative class).

The 97.4% specificity at threshold 0.3744 substantially reduces review burden, correctly identifying nearly all error-free administrations. For balanced screening, threshold 0.20 achieves 49.0% recall (detecting 152 of 310 errors) with 24.1% precision—a practical middle ground between sensitivity

<sup>1</sup>You can access the code on this link <https://github.com/msmadi/Clinical-Trial-Dosing-Error> and the dataset on HuggingFace <https://huggingface.co/datasets/sssohrab/ct-dosing-errors-benchmark>

| Metric                                                   | Score               | Note                               |
|----------------------------------------------------------|---------------------|------------------------------------|
| <b>ROC-AUC</b>                                           | <b>0.8725</b>       | Ensemble prediction (Test Dataset) |
| F1-Score                                                 | 0.2951              | Binary F1                          |
| Precision                                                | 0.3389              | Of flagged cases                   |
| Recall (Sensitivity)                                     | 0.2613              | Of actual errors                   |
| Balanced Accuracy                                        | 0.6175              | Equal class weight                 |
| Specificity                                              | 0.9737              | Of error-free cases                |
| <i>Confusion Matrix (n=6,318 samples)</i>                |                     |                                    |
| True Negatives                                           | 5,850               | 97.4% of negatives                 |
| False Positives                                          | 158                 | 2.6% false alarm rate              |
| False Negatives                                          | 229                 | 73.9% missed errors                |
| True Positives                                           | 81                  | 26.1% detected errors              |
| <i>Cross-Validation Performance (Validation Dataset)</i> |                     |                                    |
| OOV AUC                                                  | 0.8794              | Out-of-fold                        |
| Mean Fold AUC                                            | 0.8833 $\pm$ 0.0091 | 5-fold ensemble                    |

Table 5: Test Set Performance (threshold=0.3744)

and review workload. For safety-critical deployment where maximizing error detection is paramount, threshold 0.15 provides 60.3% recall (187 of 310 errors detected) at 21.4% precision—detecting the majority of errors with acceptable false positive rates for high-stakes screening applications.

## 5.2. Feature Importance Analysis

We analyze feature importance by averaging LightGBM’s gain-based metric across all 5 fold models, which measures total loss reduction attributable to each feature across all trees. Table 6 aggregates importance by category.

Word and character TF-IDF features contribute 62% of total importance despite being traditional sparse representations, challenging the assumption that dense embeddings always outperform sparse features for specialized text. Specific medical vocabulary (“reduced”, “discontinued”, “adjusted”) and morphological patterns (“discon”, “ad-jus”) carry exceptionally strong discriminative signals.

Sentence embeddings contribute 37%, with far higher average gain per feature (141.37 vs 30.40). Analysis of individual features reveals sentence embedding dimensions dominate the top-20 most important features (19 of 20), with the top 4 all being sentence embedding dimensions.

Probability scores from BiomedBERT and DeBERTa contribute only 0.06% importance. This likely results from: (1) chunk-averaging losing contextual information for long texts; and (2) single probability values providing less signal than full 768-dimensional embedding vectors.

Handcrafted features contribute 0.49%, suggesting text embedding features automatically discover the patterns we manually encoded.

## 5.3. Ablation Study

To quantify individual category contributions, we systematically remove each category and retrain using 5-fold cross-validation. Table 7 presents results.

Key findings from ablation:

**Sentence embeddings are critical:** Removing them causes the largest performance drop (2.39% degradation, from 0.879 to 0.858 AUC), confirming they capture essential semantic patterns despite contributing only 37% of total importance.

**Feature redundancy and noise:** Removing transformer scores or medical patterns slightly improves performance (-0.38%), suggesting these features introduce more noise than signal. The negative delta indicates performance improved when removed, likely due to reduced overfitting on spurious correlations.

**Strong complementarity:** Removing word/char features has minimal impact (+0.25%), showing strong redundancy with sentence embeddings when both are present.

## 5.4. Feature Efficiency Analysis

To investigate deployment efficiency and the impact of feature selection, we systematically evaluate performance using the top- $k$  most important features via 5-fold cross-validation (Table 8).

Feature selection through importance ranking reveals an intriguing pattern: using fewer features can improve performance through effective noise reduction. The optimal configuration uses 500-1000 features (14-29% of total), achieving 0.886-0.887 mean AUC—outperforming the full 3,451-feature model (0.879 AUC) by 0.7-0.8%.

This improvement demonstrates that approximately 70-85% of features in the full set contribute primarily noise rather than signal. As feature count

| Category            | Features | Total (%) | Avg Gain | Std       |
|---------------------|----------|-----------|----------|-----------|
| Word/Char Features  | 3,020    | 62.38     | 30.40    | $\pm 1.2$ |
| Sentence Embeddings | 386      | 37.07     | 141.37   | $\pm 8.5$ |
| Medical Patterns    | 43       | 0.49      | 16.84    | $\pm 0.8$ |
| Transformer Scores  | 2        | 0.06      | 43.90    | $\pm 2.1$ |

Table 6: Feature Category Importance (Averaged Across 5 Folds)

| Configuration           | Features     | Mean AUC      | Std                            | $\Delta\%$ |
|-------------------------|--------------|---------------|--------------------------------|------------|
| <b>All Features</b>     | <b>3,451</b> | <b>0.8794</b> | <b><math>\pm 0.0091</math></b> | <b>–</b>   |
| w/o Sentence Embeddings | 3,065        | 0.8584        | $\pm 0.0105$                   | +2.39%     |
| w/o Word/Char Features  | 431          | 0.8772        | $\pm 0.0087$                   | +0.25%     |
| w/o Transformer Scores  | 3,449        | 0.8827        | $\pm 0.0083$                   | -0.38%     |
| w/o Medical Patterns    | 3,408        | 0.8827        | $\pm 0.0068$                   | -0.38%     |

Table 7: Ablation Study Results (5-Fold Cross-Validation)

| K Features   | Mean AUC      | Std                            | % Baseline     |
|--------------|---------------|--------------------------------|----------------|
| 3,451 (All)  | 0.8794        | $\pm 0.0091$                   | 100.00%        |
| 3,000        | 0.8827        | $\pm 0.0085$                   | 100.38%        |
| 2,000        | 0.8837        | $\pm 0.0084$                   | 100.49%        |
| <b>1,000</b> | <b>0.8867</b> | <b><math>\pm 0.0091</math></b> | <b>100.83%</b> |
| 500          | 0.8862        | $\pm 0.0100$                   | 100.78%        |
| 200          | 0.8835        | $\pm 0.0101$                   | 100.47%        |
| 100          | 0.8802        | $\pm 0.0094$                   | 100.10%        |
| 50           | 0.8741        | $\pm 0.0105$                   | 99.40%         |
| 25           | 0.8642        | $\pm 0.0103$                   | 98.27%         |
| 10           | 0.8505        | $\pm 0.0100$                   | 96.71%         |

Table 8: Performance with Top-K Features (5-Fold CV). Bold indicates optimal performance. Feature selection improves over baseline (model with full features) through noise reduction.

decreases beyond K=500, performance gradually declines: K=200 maintains near-baseline performance (0.884, 100.47%), K=100 performs slightly above baseline (0.880, 100.10%), and K=50 shows minimal degradation (0.874, 99.40%).

The pattern of improvement-then-decline (up to k=1000) is characteristic of effective feature selection acting as regularization, removing noisy features while retaining discriminative signal. This finding has important practical implications: deployment models can be both more accurate and more efficient than the full feature set. For production deployment, we recommend K=500-1000 as optimal, providing: (1) Enhanced accuracy: +0.7% over baseline (0.886 vs 0.879 AUC), (2) Computational efficiency: 71-85% feature reduction (fewer features to compute), and (3) Robust performance: Cross-validated with  $\pm 0.009$ -0.010 standard deviation across folds.

## 5.5. Training Dynamics

Analysis of the 5-fold ensemble reveals consistent convergence behavior:

**Fold-level convergence:** Individual folds converge at 2,114 to 3,310 iterations (mean:  $2,682 \pm 485$ ) when using early stopping with 200-iteration patience. This variation reflects different training/validation partitions while maintaining robust performance.

**Cross-validation stability:** Mean fold AUC of  $0.883 \pm 0.009$  demonstrates low variance across data splits, with individual fold performance ranging from 0.869 to 0.894 (2.5% range). This consistency indicates robust learning independent of specific data partitioning.

**Generalization validation:** Out-of-fold predictions (0.8794 AUC) closely match test performance (0.8725 AUC), with only 0.7% absolute difference. This small OOF-test gap validates that the ensemble generalizes well beyond the training distribution.

**Early stopping effectiveness:** Most folds require 2,100-3,300 iterations before early stopping triggers, suggesting the problem’s complexity demands substantial boosting rounds. The patience of 200 iterations prevents premature stopping while avoiding overtraining.

The stable ensemble performance across folds and minimal OOF-test gap suggest the results will generalize to new clinical trial data from similar therapeutic areas and documentation practices.

## 6. Discussion

The findings indicate that accurate detection of dosing deviations in clinical narratives relies on combining sparse lexical features with contextual embeddings, as each captures distinct but complementary signals. Most predictive value is concentrated in a small-to-moderate subset of features (500-1000),

with the remaining features introducing more noise than signal. This suggests that efficient and interpretable models can perform competitively—and even outperform full feature sets—through judicious feature selection. However, recall–precision trade-offs highlight that such systems are best used as screening tools within human review workflows, where threshold tuning can balance safety and workload.

Our work demonstrates automated dosing error detection in clinical trial narratives through multi-modal feature engineering and ensemble learning. Key findings:

**Sparse features dominate total importance:** Word/character TF-IDF features contribute 62% of total importance despite being traditional representations, while dense embeddings contribute 37%. However, per-feature importance tells a different story: embeddings average 141.37 gain versus 30.40 for sparse features.

**Dense embeddings remain critical:** Despite lower total importance, removing sentence embeddings causes 2.4% performance degradation—the largest impact of any category. This demonstrates they capture unique semantic patterns not recoverable from lexical features alone.

**Feature selection improves performance:** The optimal 500-1000 features achieve 0.886-0.887 AUC compared to the full set’s 0.879 AUC, demonstrating that 70-85% of features contribute primarily noise. This 0.7% improvement through feature selection has important deployment implications: models can be both more accurate AND more efficient.

**Feature complementarity:** Using only embeddings (0.877 AUC) or only word/char features (ablation not directly comparable) both underperform the optimal selected subset (0.886 AUC), indicating they capture orthogonal information patterns.

**Ensemble stability:** 5-fold cross-validation achieves  $0.883 \pm 0.009$  AUC with minimal variance, and test AUC (0.873) closely matches OOF validation (0.879), demonstrating robust generalization without overfitting.

**Threshold optimization critical:** F1-optimized threshold (0.3744) yields 26.1% recall and 33.9% precision. For safety-critical deployment, threshold 0.20 achieves 49.0% recall with 24.1% precision (detecting 152/310 errors), while threshold 0.15 provides 60.3% recall at 21.4% precision (187/310 errors)—allowing flexible trade-offs based on deployment priorities.

## 7. Conclusion

We present an automated system for detecting dosing errors in clinical trial narratives, achieving 0.8725 test ROC-AUC through 5-fold ensemble

learning with comprehensive multi-modal feature engineering. Our approach combines 3,451 features spanning traditional NLP (TF-IDF, character n-grams), dense semantic embeddings (all-MiniLM-L6-v2), handcrafted medical patterns, and transformer probability scores (BiomedBERT, DeBERTa-v3).

Systematic ablation reveals: sentence embeddings cause 2.4% degradation when removed (largest impact), while word/char features show only 0.25% impact, indicating strong feature complementarity. Remarkably, feature selection improves performance: the optimal 500-1000 features achieve 0.886-0.887 AUC compared to the full set’s 0.879 AUC, demonstrating that 70-85% of features contribute primarily noise. This finding has important implications for deployment: models can be both more accurate and more efficient through judicious feature selection.

Cross-validation demonstrates robust generalization: mean fold AUC of  $0.8833 \pm 0.0091$  with minimal overfitting (OOF AUC: 0.8794, test AUC: 0.8725, gap: 0.7%). With F1-optimized threshold (0.3744), the model achieves 26.1% recall and 33.9% precision. For safety-critical deployment, threshold 0.20 achieves 49.0% recall with 24.1% precision (152/310 errors detected), while threshold 0.15 provides 60.3% recall at 21.4% precision (187/310 errors)—detecting the majority of protocol deviations with manageable false positive rates for human-review workflows.

The system advances automated protocol deviation detection, demonstrating that ensemble methods with carefully engineered features, feature selection for regularization, and hyperparameter optimization (via Optuna) remain competitive with end-to-end deep learning for specialized clinical NLP tasks with limited training data and severe class imbalance.

## 8. Bibliographical References

- Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. [Optuna: A next-generation hyperparameter optimization framework](#). In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.
- 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](#). In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78. Association for Computational Linguistics.

- 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.
- Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. [SMOTE: Synthetic minority over-sampling technique](#). *Journal of Artificial Intelligence Research*, 16:321–357.
- 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. ACM.
- Matthew M. Churpek, Richa Adhikari, and Dana P. Edelson. 2016. [The value of vital sign trends for detecting clinical deterioration on the wards](#). *Resuscitation*, 102:1–5.
- Cemal Cingi and Nuray Bayar Muluk. 2017. [Quick Guide to Good Clinical Practice](#). Springer.
- 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 (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics.
- Charles Elkan. 2001. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'01*, page 973–978, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Sohrab Ferdowsi, Félicien Hêche, Anthony Yazdani, Edward Choi, Sara Sansaloni-Pastor, and Douglas Teodoro. 2026. Overview of the ctdeb'26 shared task on predicting dosing errors in interventional clinical trials. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Mallorca, Spain.
- Sohrab Ferdowsi, Julien Knafo, Nikolay Borissov, David Vicente Alvarez, Rahul Mishra, Poorya Amini, and Douglas Teodoro. 2023. [Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study](#). *Patterns*, 4(3).
- 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.
- Rave Harpaz, William DuMouchel, Nigam H. Shah, David Madigan, Patrick Ryan, and Carol Friedman. 2012. [Novel data-mining methodologies for adverse drug event discovery and analysis](#). *Clinical Pharmacology & Therapeutics*, 91(6):1010–1021.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: Decoding-enhanced BERT with disentangled attention](#). In *International Conference on Learning Representations (ICLR)*.
- Félicien Hêche, Sohrab Ferdowsi, Anthony Yazdani, Sara Sansaloni-Pastor, and Douglas Teodoro. 2026. Early risk stratification of dosing errors in clinical trials using machine learning. *arXiv preprint arXiv:2602.22285*.
- ICH Expert Working Group. 2001. [ICH harmonised tripartite guideline: Guideline for good clinical practice E6 \(R1\)](#). International Council for Harmonisation (ICH) / regulators. Commonly accessed via regulator repositories; no Crossref DOI.
- Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu. 2019. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0). *Drug safety*, 42(1):99–111.
- Abhyuday N. Jagannatha and Hong Yu. 2016. [Structured prediction models for RNN based sequence labeling in clinical text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 856–865, Austin, Texas, USA. 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.
- Leila R. Kalankesh and Elham Monaghesh. 2024. [Utilization of ehRs for clinical trials: a systematic review](#). *BMC Medical Research Methodology*, 24(1):70.
- 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](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.



- 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.
- Fei Li, Weisong Liu, and Hong Yu. 2018. [Extraction of information related to adverse drug events from electronic health record notes: Design of an end-to-end model based on deep learning](#). *JMIR Medical Informatics*, 6(4).
- Takeo Nakayama. 2007. What are “clinical practice guidelines”? *Journal of Neurology*, 254(Suppl 5):2–7.
- Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenbourn, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. 2018. [Scalable and accurate deep learning with electronic health records](#). *npj Digital Medicine*, 1(1):18.
- 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.
- Emma Richard and Bhargava Reddy. 2021. Text classification for clinical trial operations: evaluation and comparison of natural language processing techniques. *Therapeutic Innovation & Regulatory Science*, 55(2):447–453.
- Gerard Salton and Christopher Buckley. 1988. [Term-weighting approaches in automatic text retrieval](#). *Information Processing & Management*, 24(5):513–523.
- Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. [Mayo clinical text analysis and knowledge extraction system \(cTAKES\): Architecture, component evaluation and applications](#). *Journal of the American Medical Informatics Association*, 17(5):507–513.
- Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. [Enhancing clinical concept extraction with contextual embeddings](#). *Journal of the American Medical Informatics Association*, 26(11):1297–1304.
- 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.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. [MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788.
- Yannan Yuan, Yun Mei, Shuhua Zhao, Shenglong Dai, Xiaohong Liu, Xiaojing Sun, Zhiying Fu, Liheng Zhou, Jie Ai, Liheng Ma, et al. 2024. Data flow construction and quality evaluation of electronic source data in clinical trials: pilot study based on hospital electronic medical records in china. *JMIR Medical Informatics*, 12(1):e52934.
- Andrew J. Zimolzak, Li Wei, Usman Mir, Ashish Gupta, Viralkumar Vaghani, Devika Subramanian, and Hardeep Singh. 2024. [Machine learning to enhance electronic detection of diagnostic errors](#). *JAMA Network Open*, 7(9):e2431982.

# CaresAI at CT-DEB'26: Detecting Dosing Errors In Clinical Trials Using Domain-Specific Transformer Embeddings and Classification Models

Leon Hamnett<sup>1</sup>, Favour Igwezeke<sup>1,2</sup>,  
Joseph Itopa Abubakar<sup>1</sup>, Mary Adetutu Adewunmi<sup>1,3</sup>

<sup>1</sup>CaresAI, Australia; <sup>2</sup>Faculty of Pharmaceutical Sciences, Nsukka Enugu, Nigeria;

<sup>3</sup>Menzies School of Health Research, Australia

Leon.hamnett@caresai.org, Favour.igwezeke.249461@unn.edu.ng,

Joseph.itopaa@gmail.com, Mary.adewunmi@menzies.edu.au

## Abstract

Medication errors, particularly dosing errors in clinical trials (CT), can lead to patient harm, adverse drug events and worse patient outcomes. Dosing errors are preventable, and early identification can improve trial integrity and mitigate subsequent clinical and financial burden. This study aims to detect dosing errors within CT protocols by evaluating text representations of trial information using transformer-based language models trained on biomedical corpora. CT textual data was encoded using several models, including ClinicalBERT, PubMedBERT, BioBERT, and MedCPT, and integrated with categorical features. These text embeddings were used as input to classical machine learning models and neural network architectures within an experimental framework. Performance was primarily assessed using ROC-AUC with respect to predicting dosage error. Under a logistic regression baseline, BioBERT consistently outperformed alternative encoders, achieving an ROC-AUC of 0.794, a 3.95% improvement over the ClinicalBERT baseline. Combining multiple embeddings did not yield improvements, indicating that domain alignment outweighs representational stacking. Gradient boosting models, support vector classifiers, logistic regression, and residual neural networks achieved the strongest performance for predicting dosage error, achieving ROC-AUCs: 0.821 to 0.853. Overall, the integration of domain-specific transformer embeddings with structured metadata enables discrimination of trials meeting a predefined elevated dosing error risk criterion, advancing safety monitoring and supporting informed regulatory decision-making.

**Keywords:** Medication Errors, Clinical Trials, Dosing Errors, Transformer Models, Natural Language Processing (NLP), Machine Learning

## 1. Introduction

Clinical trials provide the structured and rigorously controlled framework required to evaluate the safety and efficacy of novel pharmaceutical compounds in human populations. However, this process is both costly and uncertain. Only 14% of compounds entering clinical trials ultimately receive regulatory approval (Wong et al., 2019), despite substantial financial investment at each phase. Median costs are estimated at USD 3.4 million for Phase I, USD 8.6 million for Phase II, and USD 21.4 million for Phase III studies (Martin et al., 2017).

Beyond financial risk, trial protocols are inherently complex, involving numerous interacting variables, multiple treatment arms, and strict procedural constraints. This complexity increases the likelihood of operational and methodological errors, which may compromise statistical validity, delay development timelines, or invalidate findings entirely, resulting in wasted resources and lost therapeutic opportunity (Buyse et al., 2017; Berger and Alpers, 2009).

Medication errors are unintended failures in the drug treatment process that cause, or have the potential to cause, patient harm (European

Medicines Agency, 2026). These errors can include: incorrect timing of doses, incorrect dosage, additional unintended doses, or inappropriate administration rates (Tariq et al., 2024), as well as errors in randomisation, masking, and broader experimental design (Berger and Alpers, 2009). Medication errors remain one of the most frequently occurring yet preventable causes of adverse patient outcomes within clinical settings (Tariq et al., 2024). The same medication errors that occur in clinical settings can also arise during clinical trials. Medication errors during clinical trials can lead to reduced efficacy of the trial, serious toxicity, compromised patient safety, adverse drug events, and can also pose a threat to internal validity and interpretability of clinical trial outcomes (European Medicines Agency, 2026).

Given the substantial consequences of erroneous approval decisions related to new compounds, rigorous monitoring and statistically robust evaluation are essential during the entire clinical trial process. Regulatory bodies such as the Food and Drug Administration (FDA) and European Medicines Agency (EMA), emphasise the importance of proactively identifying, monitoring and mitigating dosing errors during clinical trials due to

their potential to contribute to a public health burden (Stewart, 2021). In addition, timely identification of dosing errors in clinical trials can reduce both financial costs and patient risk by facilitating earlier correction or termination of ineffective studies (Rim et al., 2024).

However, clinical trial documentation often consists of large volumes of unstructured textual data, making the extraction of relevant methodological details time-consuming and challenging (Chen et al., 2020). This complicates verification of trial integrity, particularly with respect to correct dosing and administration schedules across study arms and cohorts. The predominance of unstructured textual documentation within clinical trial reports and protocols, highlights the need for automated methods capable of systematically extracting and verifying critical methodological details.

In the context of the medical field and more specifically when evaluating clinical reports, Natural Language Processing (NLP) provides a scalable approach for identifying and structuring relevant information embedded within narrative clinical text. Natural language processing is a subfield of computer science and artificial intelligence, which allows computers and digital devices to recognise and understand text and speech data (IBM, 2026). Beyond information extraction from text, NLP techniques also support downstream analytical tasks such as diagnosis explanation, disease progression modelling, and treatment effectiveness assessment. Recent advances in machine learning have led to the development of increasingly sophisticated NLP algorithms with broad applications in information extraction and summarisation (Khurana et al., 2023). These developments within the field of NLP are particularly relevant to medical documentation, where clinical notes and reports generated from medical practise and research typically take the form of large volumes of unstructured text. NLP methods have already been applied within the medical field for some use-cases, including: extracting dosage information and administration schedules (Lu et al., 2016; McNeer et al., 2019), identifying clinical and treatment events (Bitterman et al., 2023), and detecting adverse drug events and reactions (Murphy et al., 2023; Naderian et al., 2024). Recent research has also investigated specific applications of these NLP methodologies within the context of clinical trials, including: automating report drafting, extracting drug-to-drug interactions (DDI) and optimising protocol design in CTs (Chopard et al., 2021; Kuo et al., 2025; Lee et al., 2024; Zhu et al., 2020).

The rapid proliferation of transformer-based language models has significantly advanced NLP research (Kalyan et al., 2021; Patwardhan et al.,

2023). These models leverage self-supervised learning and attention mechanisms to learn feature-rich, general-purpose representations from large volumes of unlabelled text corpora. As a consequence of extensive pre-training on a large amount of data, only relatively small amounts of task-specific data are required to achieve strong performance on downstream tasks, since the pre-trained parameters within each model already encode substantial linguistic and semantic structure derived from the training corpora.

During pre-training, transformer architectures learn contextualised representations of input text, capturing semantic and syntactic relationships within fixed-length numerical embeddings. Owing to their success in general-domain NLP, these models have increasingly been applied in the medical domain across tasks such as question answering, named entity recognition (NER), text summarisation, classification, and sentiment analysis (Nerella et al., 2024; Cho et al., 2024).

More recently, research has investigated whether domain-specific pre-training on biomedical corpora, such as PubMed abstracts and Mayo Clinic electronic health records, can further improve performance in a medical context (Wang et al., 2018). Empirical findings suggest that models pre-trained predominantly on medical text corpora achieve superior results on clinical tasks (Wang et al., 2018). Exposure to specialised clinical terminology and domain-specific vocabulary enhances representational precision, while repeated exposure to clinical narratives allows models to implicitly capture the structural patterns of medical documentation. Moreover, domain-aligned pre-training reduces distributional shift between pre-training data and downstream data, improving robustness and sample efficiency. This is particularly important in clinical contexts, where annotated datasets are scarce, costly to produce, or limited in scope.

The strong performance of pre-trained transformer models in NER tasks enables reliable extraction of key clinical entities, including diseases, medications, procedures, and intervention attributes. This capability is especially relevant for clinical trial analysis, where experimental design elements, such as study arms, intervention types, and dosing schedules, are frequently recorded as unstructured text. There has already been some success in applying pre-trained transformer models within the context of clinical trials, including tasks related to: extracting outcome and arm information (Whitton and Hunter, 2023), distinguishing adverse event reporting (Chopard et al., 2021), and extracting drug-drug interactions (Zhu et al., 2020). However, despite these recent advances, access to large, openly available clinical trials datasets remains limited, restricting large-scale experimentation and

hypothesis testing (Navar et al., 2016). Recent initiatives such as TrialBench have sought to address this gap through multi-modal dataset construction and standardised baselines for clinical trial prediction tasks (Chen et al., 2025).

Consequently, transformer-based models provide a scalable framework for structuring and validating critical methodological information embedded within clinical trial documentation. As part of this research project, the authors investigated the suitability of various transformer-based models which have been pre-trained on biomedical text corpora, in order to encode the textual information related to each clinical trial, and allow machine learning models to make predictions on whether a specific clinical trial contains a dosage error within its approach.

## 2. Methodology

### 2.1. Dataset and Problem Formulation

As part of the CT-DEB'26 challenge (Detecting Dosing Errors from Clinical Trials) (Ferdowsi et al., 2026), a dataset was generated (Hêche et al., 2026), containing structured and unstructured information for a large number of different clinical trials (written in the English language), collected from the ClinicalTrials.gov site (U.S. National Library of Medicine, 2026). Each row corresponded to a single clinical trial study and included long-form textual fields as well as several numerical and categorical variables related to each specific trial. As part of the dataset construction process, a binary label was generated: a label of 1 was assigned if the lower bound of the 95% Wilson confidence interval for the dosing error rate exceeded 0.01% (Ferdowsi et al., 2026; Hêche et al., 2026).

The objective of this research was to compare several pre-trained biomedical models for encoding text features, and to develop and evaluate machine learning models (both classical and neural networks) capable of predicting which clinical trials met this elevated dosing error risk criterion. The study systematically evaluated the impact of different feature subsets, alternative input representations of textual data, and varying model architectures on predictive performance.

Given the presence of substantial unstructured text, it was necessary to transform textual data into machine-readable representations. To this end, several pre-trained transformer-based models were evaluated, including ClinicalBERT (Huang et al., 2020), BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021), and MedCPT (Jin et al., 2023). These models, pre-trained on large-scale biomedical corpora, can generate contextualised vector representations of input text

that capture domain-specific semantic information. The resulting embeddings were used as input features for downstream machine learning classifiers, including both traditional machine learning models, as well as more advanced neural networks. A controlled experimental framework was applied throughout the research process, modifying a single variable at a time while holding others constant to isolate individual effects of experimental changes. Standard machine learning practice was followed by using a separate subset of data to analyse the predictive performance of each model, to prevent bias by showing the model data it has already seen before during the training process.

The dataset was provided in two separate subsets: a training dataset of 29,478 data points and a hold-out validation dataset of 6,316 data points. Each dataset contained the following columns:

phases, enrollmentCount, allocation, interventionModel, primaryPurpose, masking, healthyVolunteers, sex, oversightHasDmc, briefSummary, detailedDescription, conditions, conditionsKeywords, protocolPdfText, numArms, armDescriptions, armGroupTypes, numInterventions, interventionTypes, interventionDescriptions, interventionNames, numLocations, locationDetails, target, nctid.

A detailed description of the dataset variables can be found in Appendix A.1. The dataset contained a mixture of categorical, numerical, and free-form text variables, each related to a specific clinical trial.

In the final modelling pipeline, textual features (briefSummary, detailedDescription, armDescriptions, and interventionDescriptions) were encoded using BioBERT embeddings. Structured categorical variables, including allocation method, intervention model, primary purpose, oversightHasDmc, and trial phase, were one-hot encoded and concatenated with the text embeddings. Continuous numerical variables such as enrollmentCount, numArms, and numLocations were not included in the final models. The numArms variable demonstrated a weak statistical association with the target variable, while enrollmentCount and numLocations were not considered likely to provide meaningful predictive signal beyond the structured and textual features already incorporated.

The protocolPdfText field was not incorporated into the modelling pipeline due to substantial data sparsity and extreme variability in length. In the training set, 81.3% of records (23,963 of 29,478) contained no extracted PDF text, whereas only 3.2% of validation records were empty. This



imbalance raised concerns that the model could implicitly learn the presence of non-empty PDF text as a proxy for dataset partition, thereby introducing distributional bias. Furthermore, when data was available within the protocolPdfText field, the extracted protocol text was extremely long (training max: 11,298,927 characters; validation max: 6,424,120 characters), with highly skewed distributions and very large standard deviations. Encoding such long documents within a fixed transformer context window would require aggressive truncation, potentially discarding the majority of the content and introducing inconsistent representation quality across samples. Given the high proportion of missing entries in training data, the substantial computational burden of processing very long documents, and the risk of introducing noise through heavily truncated representations, this feature was excluded from the final experiments.

As the primary objective of this study was to evaluate the comparative effectiveness of pre-trained biomedical language models for encoding unstructured clinical trial documentation, the modelling framework was designed to isolate the contribution of textual representations. Accordingly, only categorical structured features were included alongside the embeddings, and continuous numerical variables were excluded.

## 2.2. Overview of Experimental Investigations

An iterative and controlled experimental framework was adopted. At each stage, only one variable or modelling component was modified while all other elements were held constant. This enabled the isolation of the effect of individual design decisions on predictive performance.

The experimental process was conducted in two major phases:

Phase 1: Iteration of text encoding methods and feature selection

Phase 2: Iteration of model architectures and training processes

### 2.2.1. Phase 1: Text Encoding and Importance of Features

**Categorical Feature Analysis:** Initial experiments employed a logistic regression classifier as a baseline model. Using this fixed architecture, the study evaluated the incremental impact of incorporating one-hot encoded categorical variables. Associations between categorical features and the target variable were assessed using Chi-squared tests and Cramér's V statistics. Importantly, these statistical analyses were conducted exclusively on the training data to prevent information leakage, ensuring that the validation set remained fully

unseen during both feature selection and model evaluation to give unbiased reporting of the models performance on unseen data.

An ablation study was conducted by incrementally adding individual features to measure their contribution to predictive performance. Categorical features were selected to be used in the following stages, primarily based on statistical significance in relation to the target feature.

**Text Encoding with Pre-trained Transformer Models:** Given the presence of substantial unstructured text, multiple pre-trained biomedical transformer encoders were evaluated to generate contextual embeddings for the following textual columns: "Brief summary", "Detailed description", "Arm descriptions", "Intervention descriptions". The following encoders were compared: ClinicalBERT, PubMedBERT, BioBERT, MedCPT, and also combinations of multiple encoders. A short description of each transformer model is given below:

- **BioBERT** is a domain-specific language model pre-trained on large-scale biomedical literature, including PubMed abstracts and PubMed Central full-text articles, enabling it to capture biomedical terminology and contextual relationships in clinical text (Lee et al., 2020).
- **ClinicalBERT** is a BERT-based model pre-trained on clinical notes from intensive care unit records in the MIMIC-III database, making it well suited for clinical narrative understanding tasks (Huang et al., 2020).
- **MedCPT** is a contrastive learning-based biomedical language model trained on 255 million user click logs from PubMed, designed to produce high-quality semantic embeddings for zero-shot biomedical information retrieval (Jin et al., 2023).
- **PubMedBERT** is a language model pre-trained from scratch exclusively on PubMed abstracts, without general-domain pre-training, allowing it to develop vocabulary and representations specifically optimised for biomedical text (Gu et al., 2021).

Each pre-trained encoder generated fixed-length vector representations for the textual fields. A maximum sequence length of 256 tokens was selected as a practical compromise between contextual coverage and computational efficiency. Token-level analysis using the BioBERT tokenizer (see Appendix A.2) indicates that, under this constraint, the proportion of truncated entries was 8.65% for briefSummary, 31.81% for detailedDescription, 7.79% for armDescriptions, and 4.03% for interventionDescriptions. Notably, the detailedDescription field contained a substantial



proportion of empty entries 39.4% in the training set (11604 of 29478 rows) and 34.6% in the validation set (2187 of 6316 rows), which reduces the effective truncation impact for that field. Overall, more than 90% of entries in briefSummary, armDescriptions, and interventionDescriptions were fully preserved within the 256-token limit; detailedDescription experienced higher truncation (31.81%), though a substantial proportion of entries were empty. A uniform maximum token length was applied across all textual fields to ensure consistency within the encoding pipeline. While increasing the context window may allow additional long-range information to be captured, particularly for the detailedDescription field, this would substantially increase computational cost and memory requirements. The selected limit therefore reflects a trade-off between representational completeness and practical feasibility.

Sentence-level embeddings were obtained using mean pooling over the token-level representations produced by each model. These embeddings were then used as input features for a downstream logistic regression classifier, which learned to map the numerical vector representations to the binary dosing error label. Comparative experiments evaluated the impact of different pre-trained encoders on validation ROC-AUC performance. BioBERT achieved the strongest results and was therefore selected as the primary text encoder for subsequent experiments.

### **2.2.2. Phase 2: Iteration of Model Architectures and Training Processes**

The final set of inputs features was selected based on statistical analysis and domain knowledge, by combining BioBERT embeddings for all text fields and one-hot encoded categorical features (specifically: intervention model, allocation type, oversight, primary purpose, phase). These features were features that gave the model the most discriminative predictive power in determining if the probability of a dosage error having occurred within a specific clinical trial exceeded the defined threshold.

As the dataset exhibited substantial class imbalance (approximately 95.4% of training records labelled as no dosage error - 28,126 of 29,478 data-points), it was necessary to investigate methodologies for mitigating this large imbalance. A series of experiments evaluated the impact of varying class weights during training. By varying the class weights, we can instruct a machine learning model to place greater importance on examples from a certain class during the training process. These experiments consisted of the following stages: no weighting (all examples assigned equal importance), fixed weighting ratios

(1:5, 1:10, 1:20) toward the minority class, and automatically computed class weights based on the proportional amount of each element within the dataset (around 0.5:11 weighting, for 0 and 1 dosage error variables, respectively).

Once the final set of input features was fixed, and an optimal class weighting had been identified, several experiments were undertaken to vary the type of machine learning model that was used to make predictions, and thus change the manner in which the model is learning from the input data. These models included both classical machine learning models and more advanced neural network architectures. The classical machine learning models that were examined were as follows: Logistic Regression, Linear Support Vector Classifier, Support Vector Classifier, K-Nearest Neighbours Classifier, Naïve Bayes, Decision Trees, XGBoost and LightGBM. Due to the high dimensionality of transformer embeddings, some classical models experienced convergence or computational limitations. Comparative evaluation identified Logistic Regression, SVC variants, as well as Gradient Boosting models as the strongest classical baselines. The top performing models had hyperparameter optimisation performed using GridSearch to identify the combination of model parameters that gave the best predictive power.

Following the evaluation of the classical machine learning models, multiple neural network architectures were explored: simple multi-layer perceptron (MLP), wide multi-layer perceptron, residual neural network (with skip connections), text-tower architecture (separate processing streams for text and categorical features, merged at a later part of the neural network). Specific implementation details of the various neural network architectures can be found in [Appendix A.6](#).

Performance differences between neural network architectures was analysed using consistent feature representations and with other neural network training parameters fixed (e.g. loss, optimiser, learning rate, training epochs). During the later stages of investigation into neural network models, once the best performing architecture was identified, the training process itself was further optimised. This included the following methodologies: early stopping based on validation ROC-AUC, learning rate scheduling (ReduceLROnPlateau, OneCycleLR), loss function comparison (binary cross-entropy vs focal loss), and model check-pointing to retain best parameters against the validation data. These experiments aimed to optimise convergence behaviour during model training and reduce over-fitting to the training data. All models were evaluated on held-out validation data to assess generalisation performance on unseen data. The primary evaluation metric

was ROC-AUC, with secondary metrics used to characterise performance under class imbalance conditions.

Extensive hyperparameter optimisation of the neural network models was not performed. Instead, the architectures were constructed to represent common and practically relevant structural variations, such as wide networks, residual connections, and gated text towers, rather than to exhaustively tune each configuration for maximal performance. Although more comprehensive optimisation could potentially improve absolute predictive performance, this was beyond the scope of the present study. The reported results therefore reflect controlled and representative configurations designed to enable a fair and consistent comparison across modelling approaches.

### 2.2.3. Phase 3: Final Modelling

Following identification of the strongest-performing classical models and neural network architectures, a final evaluation stage was conducted to obtain a more robust estimate of predictive performance. The original training and validation datasets were combined, and stratified k-fold ( $k=5$ ) cross-validation was applied to create unique data subsets while preserving the overall proportion of positive and negative cases within each fold.

In each iteration, the model was trained on four folds and evaluated on the remaining fold, producing five independent sets of performance metrics. Averaging these results provides a more reliable estimate of generalisation performance by reducing dependence on any single train-validation partition. Stratification further ensures that class imbalance is consistently represented across folds, thereby limiting variability caused by uneven class distributions and yielding a more stable and unbiased assessment of model performance.

## 3. Discussion of Results

### 3.1. Results for Encoding of Text Columns

BioBERT demonstrated the strongest individual encoder performance, achieving a ROC-AUC of 0.794 against the validation set under logistic regression, compared to 0.775 for PubMedBERT, 0.789 for MedCPT, and 0.764 for ClinicalBERT (see Figure 1). This is consistent with the design of BioBERT, which was pre-trained on large-scale biomedical literature, including PubMed abstracts and PMC full-text articles, enabling it to capture domain-specific linguistic patterns relevant to clinical trial intervention descriptions. As shown in Figure 2, when compared to the ClinicalBERT baseline encoding method, BioBERT

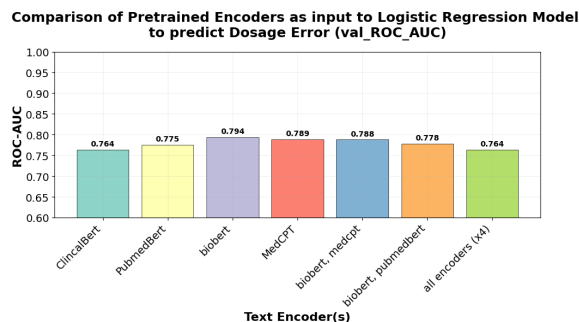


Figure 1: Comparison of pretrained encoder models (absolute ROC-AUC)

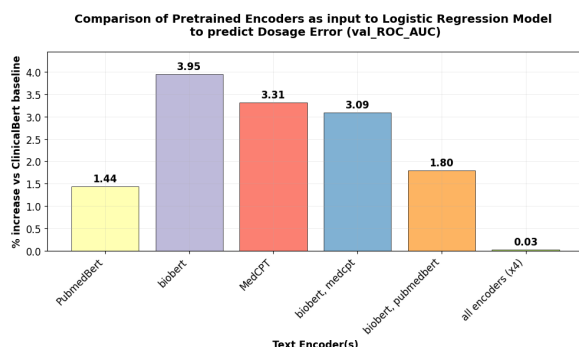


Figure 2: Comparison of pretrained encoder models (% ROC-AUC improvement over ClinicalBERT baseline)

shows close to a 4% improvement in ROC-AUC. ClinicalBERT, while pre-trained on clinical notes from intensive care unit records, may be less suited to the structured protocol language characteristic of ClinicalTrials.gov records (Alsentzer et al., 2019). The more advanced model architectures, including XGBoost, LightGBM, and residual neural networks, achieved stronger predictive performance when using BioBERT embeddings, with ROC-AUC values ranging from 0.821 to 0.832. These results suggest an enhanced capacity to learn from the BioBERT vector representations and improve discriminative ability in identifying trials meeting the predefined elevated dosing error risk threshold.

Contrary to the initial expectation that combining multiple encoders would yield additive gains, results under logistic regression showed that the combined four-encoder representation (ROC-AUC: 0.764) under-performed BioBERT alone, showed minimal difference in ROC-AUC compared to the ClinicalBERT baseline. This suggests that when the downstream model lacks sufficient capacity to exploit high-dimensional concatenated representations, additional encoders may introduce noise rather than complementary discriminative ability. It is possible that the more advanced gradient

boosting models and neural networks may be better able to utilise text encodings from multiple encoders at once to boost predictive ability, but this subject was not rigorously examined as part of this study.

### 3.2. Results for Importance of Categorical Features

The inclusion of one-hot encoded categorical features consistently improved model performance. A statistical analysis combining Chi-squared and Cramer's V, identified *interventionModel* ( $\chi^2=219.13$ ,  $p<0.001$ ), *allocation* ( $\chi^2=90.88$ ,  $p<0.001$ ), and *oversightHasDmc* ( $\chi^2=76.83$ ,  $p<0.001$ ) as the most statistically significant predictors of dosing error occurrence. From a clinical perspective, this is intuitive: crossover and factorial intervention designs involve more complex dosing schedules and transitions between treatment arms, increasing the procedural risk of dosing errors compared to simpler parallel or single-group designs (England et al., 2020). Similarly, the absence of a Data Monitoring Committee may reflect reduced oversight infrastructure, as shown in prior work to be associated with higher rates of protocol deviation and adverse events in clinical trials (Stewart, 2021).

Although the categorical features 'primary purpose' and 'trial phase' did not meet the threshold for statistical significance with regard to impacting the dosage error target feature, domain knowledge supported their inclusion. Their addition produced improvements in several model configurations and did not appear to actively reduce predictive ability, and as such were included in the subsequent experiments. This reflects the broader challenge of feature selection in clinical prediction tasks, where statistical significance does not always align with the predictive ability of machine learning models.

### 3.3. Results for Class Weighting

A significant methodological challenge in this study was the pronounced class imbalance within the dataset, with approximately 95.4% of training records labelled as 0 (28,126 of 29,478), indicating that the lower bound of the 95% Wilson confidence interval for the dosing error rate did not exceed the predefined threshold for those clinical trials. This distribution reflects the real-world rarity of formally reported dosing errors in clinical trial registries, where medication errors are known to be systematically under-reported (Delavoipière et al., 2021). Moreover, using ROC-AUC as an evaluation metric can be misleading in highly imbalanced applications, as strong performance may still coincide with poor minority class detection (Imani et al., 2026). To address this, additional evaluation metrics, including precision, recall, and

balanced accuracy, were incorporated to ensure that models were meaningfully identifying cases where dosing error likelihood surpassed the defined threshold, rather than defaulting to majority-class predictions.

When using simpler machine learning models such as logistic regression, varying the class weights for the minority class had no meaningful impact on logistic regression performance, with ROC-AUC against the validation set remaining within a narrow range (0.791–0.798) across all weighting strategies (mean: 0.795, standard deviation of 0.003). This suggests that re-weighting did not materially alter ranking-based discriminative performance with respect to ROC-AUC. Alternative approaches, such as upsampling the minority class, may potentially improve sensitivity; however, this was not explored in the present study. In medical contexts, synthetic or repeated sampling strategies must be applied cautiously, as they may introduce over-fitting, distort clinically meaningful distributions, or amplify noise in rare but heterogeneous cases.

In contrast to logistic regression, the neural network demonstrated improved ability to predict dosage error when class weighting was applied, with ROC-AUC increasing from 0.806 (no weighting) to 0.838 using automatically computed class weights based on the proportion of examples of each class within the training data. This indicates that the neural architecture was more responsive to imbalance-aware training, improving its ability to detect minority class instances (dosing error probability surpassed defined threshold) without sacrificing overall discriminative performance (mean: 0.835, standard deviation of 0.002, across experiments where class weighting was applied).

### 3.4. Results for Comparison of Classical ML models

As shown in Figure 3, the strongest discriminative performance on the validation set, as measured by ROC-AUC, was achieved by LightGBM, XGBoost, the calibrated support vector classifier (SVC), and logistic regression. These models consistently outperformed simpler methods such as k-nearest neighbours and decision trees, indicating that either well-regularised linear decision boundaries or gradient-boosted ensemble approaches are particularly well suited to the high-dimensional feature space produced by transformer-based text embeddings combined with categorical variables.

Following light hyperparameter optimisation, LightGBM achieved a validation ROC-AUC of 0.831 using `num_leaves = 31`, `max_depth = 8`, `learning_rate = 0.05`, and `min_child_samples = 50`. Logistic regression

achieved a validation ROC-AUC of 0.812 with an L2 penalty and regularisation strength  $C = 0.1$  and 'balanced' class weighting. The calibrated SVC achieved a cross-validated ROC-AUC of 0.831 with  $C = 1$ ,  $\gamma = 0.001$ , and an RBF kernel. These findings suggest that both ensemble-based methods and regularised margin-based classifiers are capable of effectively leveraging the semantic representations generated by domain-specific transformer encoders.

Notably, both tree-based boosting methods and calibrated linear models achieved comparable predictive performance, indicating that substantial gains can be achieved without necessarily increasing architectural complexity. From a practical perspective, logistic regression offers an additional advantage in terms of interpretability. The learned coefficients provide a direct estimate of feature contribution to the predicted probability of a dosing error, enabling transparent inspection of influential variables. When combined with L1 regularisation, or elastic net regularisation with a non-zero L1 ratio, the model promotes sparsity by shrinking less informative coefficients to zero. This facilitates feature selection, reduces model complexity, and enhances interpretability, an important consideration in clinical and regulatory settings where model transparency and justification of decisions are critical.

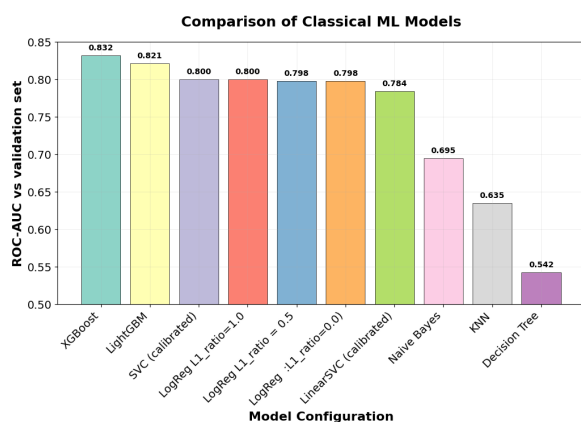


Figure 3: Comparison of Classical ML Models

### 3.5. Results for Comparison of Neural Network models

A series of neural network architectures was evaluated to assess the impact of structural design on predictive performance. A simple multi-layer perceptron achieved a validation ROC-AUC of 0.829, providing a strong baseline (see Figure 4). Expanding the architecture into a wider network did not yield improvements, resulting in a lower ROC-AUC of 0.814, suggesting that

increasing parameter count alone did not enhance generalisation. Introducing residual connections led to the strongest performance, with a ROC-AUC of 0.831, indicating that skip connections may facilitate more effective gradient flow and feature utilisation in the high-dimensional embedding space. Additionally, a text tower architecture, designed to process textual embeddings separately from categorical features before merging, achieved ROC-AUC scores of 0.820 (with gating) and 0.822 (without gating). These results suggest that while architectural modifications influence performance, residual connections provided the most consistent gains, whereas increased width or gated fusion mechanisms within the neural networks did not confer additional benefit.

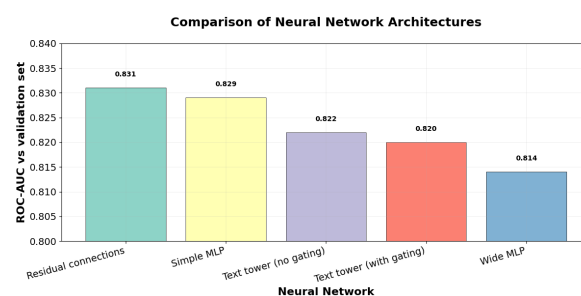


Figure 4: Comparison of Neural Network Models

### 3.6. Results after Cross-validation

Following model selection and hyperparameter optimisation via GridSearch (only applied for classical ML models, specific GridSearch parameters can be seen in Appendix A.4), a final evaluation was conducted using stratified k-fold cross-validation (k=5) across the combined training and validation datasets. Full fold-level results are reported in Appendix A.5.

The light gradient boosting model, calibrated SVC model, and the text-tower neural network, all demonstrated stronger and more stable discriminative ability across folds. For these models, performance under stratified five-fold cross-validation was consistent with the previous train-validation, indicating stability across sampling splits. The calibrated SVC achieved a mean ROC-AUC of  $0.841 (\pm 0.007)$ , while LightGBM achieved  $0.836 (\pm 0.010)$ , and the text-tower neural network achieved  $0.831 (\pm 0.012)$ . The relatively low standard deviations across folds indicate robustness to sampling variability, suggesting that these models generalise consistently across different subsets of the dataset. Although the calibrated SVC achieved the highest ROC-AUC, its recall at a default decision threshold of 0.5 remained low,



reflecting conservative positive class predictions under this operating point.

When comparing performance across the stratified k-fold ( $k=5$ ) cross-validation data splits against the results when trained with the provided train-validation split, the logistic regression models showed a reduction in mean ROC-AUC under cross-validation. Logistic regression achieved a mean ROC-AUC of  $0.732 (\pm 0.006)$ . This decline suggests that earlier performance against the initial training and validation sets may have been biased due to variations in data partitioning. In particular, the reduced performance indicates that this approach may rely more heavily on fold-specific structure within the data, resulting in lower generalisation when evaluated across multiple stratified splits.

All classification metrics were computed using a fixed probability threshold of 0.5 to ensure consistency across models. Given the pronounced class imbalance in the dataset, this threshold may not be optimal for maximising clinically meaningful trade-offs between precision and recall. In practice, threshold tuning based on application-specific priorities (e.g. prioritising sensitivity for early risk detection) could further improve precision, recall, and balanced accuracy without altering the underlying ranking performance as measured by ROC-AUC.

Overall, the cross-validation results provide a more reliable estimate of model generalisation performance and highlight the relative robustness of the calibrated SVC and text-based neural architectures for this task.

### 3.7. Summary of Findings

This study evaluated transformer-based biomedical text embeddings combined with structured categorical features for predicting clinical trial protocols in which the probability of a dosing error exceeded a defined threshold. Model performance was assessed primarily using ROC-AUC, with secondary metrics including F1-macro, precision, recall, and balanced accuracy. Across all experiments, BioBERT consistently outperformed alternative biomedical language models when used as a stand-alone encoder. Furthermore, combining multiple pretrained encoders did not produce additive gains, suggesting that domain alignment was more important than representational stacking. Incorporating structured categorical trial design variables alongside text embeddings improved predictive performance across model classes.

Among the evaluated approaches, residual neural networks, calibrated support vector classifiers, and light gradient achieved the strongest and most stable performance under cross-validation, while gradient boosting models achieved competitive

results with lower computational cost. Explicit handling of class imbalance was necessary to obtain meaningful recall of the minority class. These findings highlight the importance of domain-specific pretraining and careful model selection in highly imbalanced clinical prediction tasks.

## 4. Conclusion

This work demonstrates that domain-specific transformer embeddings, particularly BioBERT, provide highly effective representations for modelling unstructured clinical trial documentation. When integrated with structured trial metadata, these embeddings enable reliable discrimination of trials in which the estimated probability of dosing error exceeds a predefined threshold.

The results suggest that encoder-domain alignment plays a greater role in predictive performance than increasing architectural complexity or combining multiple pretrained representations. While neural architectures achieved strong performance, gradient boosting methods offered competitive discriminative ability with greater interpretability and lower computational cost, making them attractive candidates for deployment in regulatory or governance settings where transparency is essential.

Future research may explore threshold optimisation strategies tailored to specific clinical priorities, the application of model explainability techniques to identify high-risk protocol elements, and the extension of this framework to multi-class prediction of dosing error types. Incorporating additional data sources, such as adverse event reports or electronic health records, may further strengthen real-world applicability and support proactive safety monitoring.



## 5. Limitations

Firstly, compute constraints prevented the evaluation of language models with larger amounts of parameters, such as BioGPT (Luo et al., 2022) and SciFive (Phan et al., 2021). Encodings from these models may have produced stronger semantic representations of clinical trial protocol text. Secondly, the binary target variable conflates dosing errors of varying severity and type into a single label, limiting the granularity of predictions. Future work could explore multi-class formulations that distinguish between error types such as incorrect dose, incorrect frequency, and omission errors, as taxonomised in prior work (England et al., 2020). Thirdly, the dataset is derived exclusively from the ClinicalTrials.gov registry, which captures structured protocol information but may not reflect the full complexity of dosing procedures described in supplementary protocol documents or internal trial management systems. As noted in prior work (Navar et al., 2016), registry data is subject to selective reporting and inconsistent documentation practices.

Finally, while feature importance scores can be extracted from XGBoost and LightGBM, interpreting which specific textual patterns in protocol descriptions drive predictions remains an open challenge that warrants further investigation using explainability techniques such as SHAP values or attention-based visualisation. In addition, the neural network models provide difficulty in explainability of results due to the 'black-box' nature of the neural network's internal learning processes.

## 6. Ethics Statement and Code/Data Availability

This study utilised publicly available clinical trial registry data and secondary benchmark healthcare datasets via the Hugging Face online platform and was subject to the original data providers' terms and conditions. Due to licensing and governance restrictions, raw clinical trial and healthcare datasets are not redistributed by the authors. Additionally, the datasets were anonymised in accordance with their respective usage policies. As such, the research did not require additional ethical approval. The study was conducted in accordance with relevant data governance standards and responsible AI research practices.

To facilitate open research and transparency, the code for this study is publicly available via the [GitHub repository](#).

## 7. Bibliographical 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.
- Victor W. Berger and Sharon Y. Alperson. 2009. [A general framework for the evaluation of clinical trial quality](#). *Reviews on Recent Clinical Trials*, 4(2):79–88.
- Danielle S. Bitterman, Eli Goldner, Sean Finan, David Harris, Eric B. Durbin, Harry Hochheiser, Jeremy L. Warner, Raymond H. Mak, Timothy Miller, and Guergana K. Savova. 2023. [An end-to-end natural language processing system for automatically extracting radiation therapy events from clinical texts](#). *International Journal of Radiation Oncology\*Biophysics*, 117(1):262–273.
- Marc Buyse, Pierre Squifflet, Evelyne Coart, Emilie Quinaux, Cornelis J. A. Punt, and Everardo D. Saad. 2017. [The impact of data errors on the outcome of randomized clinical trials](#). *Clinical Trials*, 14(5):499–506.
- J. Chen, Y. Hu, M. Cai, Y. Lu, Y. Wang, X. Cao, M. Lin, H. Xu, J. Wu, and X. Cao. 2025. Trialbench: Multi-modal ai-ready datasets for clinical trial prediction. *Scientific Data*, 12(1):1564.
- Xiang Chen, Haoran Xie, Gong Cheng, Lai K. M. Poon, Min Leng, and Fei L. Wang. 2020. [Trends and features of the applications of natural language processing techniques for clinical trials text analysis](#). *Applied Sciences*, 10(6):2157.
- Hyeon Nam Cho, Tae Jin Jun, Young Ho Kim, Hyun Kang, Inseok Ahn, Hyuk Gwon, Yejin Kim, Jiwon Seo, Hyun Choi, Minji Kim, Jiwon Han, Gyeong Kee, Seung Park, and Seung Ko. 2024. [Task-specific transformer-based language models in health care: Scoping review](#). *JMIR Medical Informatics*, 12:e49724.
- Daniel Chopard, Matthias S. Treder, Paul Corcoran, Nadeem Ahmed, Chris Johnson, Monika Busse, and Irena Spasić. 2021. [Text mining of adverse events in clinical trials: Deep learning approach](#). *JMIR Medical Informatics*, 9(12):e28632.
- Emilie Delavoipière, Céline Fourage, Matthieu Macro, et al. 2021. [\[medication errors reporting](#)

- in drug clinical trials: Role of the clinical research pharmacist?]. *Thérapie*, 76(6):735–742.
- Eleanor England, Charles D. Deakin, Jerry P. Nolan, Rajesh Lall, Tom Quinn, Simon Gates, Jonathan Miller, Louise O'Shea, Hannah Pocock, and Nicola Rees. 2020. Patient safety incidents and medication errors during a clinical trial: experience from a pre-hospital randomized controlled trial of emergency medication administration. *European Journal of Clinical Pharmacology*, 76(10):1355–1362.
- European Medicines Agency. 2026. [Medication errors](#). Overview of EMA's role in pharmacovigilance and medication error reporting and prevention in the European Union.
- Sohrab Ferdowsi, Félicien Hêche, Anthony Yazdani, Edward Choi, Sara Sansaloni-Pastor, and Douglas Teodoro. 2026. Overview of the ct-deb'26 shared task on predicting dosing errors in interventional clinical trials. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Mallorca, Spain.
- 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).
- Félicien Hêche, Sohrab Ferdowsi, Anthony Yazdani, Sara Sansaloni-Pastor, and Douglas Teodoro. 2026. Early risk stratification of dosing errors in clinical trials using machine learning. *arXiv preprint arXiv:2602.22285*.
- Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. [Clinicalbert: Modeling clinical notes and predicting hospital readmission](#).
- IBM. 2026. [What is natural language processing \(nlp\)?](#)
- M. Imani, M. Joudaki, A. Bagheri, and H. R. Arabnia. 2026. [Why ROC-AUC is misleading for highly imbalanced data: In-depth evaluation of MCC, F2-score, H-measure, and AUC-based metrics across diverse classifiers](#). *Technologies*, 14(1):54.
- 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.
- Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. [Ammus : A survey of transformer-based pretrained models in natural language processing](#).
- Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. [Natural language processing: state of the art, current trends and challenges](#). *Multimedia Tools and Applications*, 82:3713–3744.
- S.-M. Kuo, S.-K. Tai, H.-Y. Lin, and R.-C. Chen. 2025. [Automated clinical trial data analysis and report generation by integrating retrieval-augmented generation \(RAG\) and large language model \(LLM\) technologies](#). *AI*, 6(8):188.
- 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.
- K. Lee, Z. Liu, Y. Mai, T. Jun, M. Ma, T. Wang, L. Ai, E. Calay, W. Oh, G. Stolovitzky, E. Schadt, and X. Wang. 2024. [Optimizing clinical trial eligibility design using natural language processing models and real-world data: Algorithm development and validation](#). *JMIR AI*, 3:e50800.
- Ching-Chi Lu, Jinyuan Leng, George W. Cannon, Xiaoming Zhou, Matthias Egger, Brett South, Zachary Burningham, Qing Zeng, and Brian C. Sauer. 2016. [The use of natural language processing on narrative medication schedules to compute average weekly dose](#). *Pharmacoepidemiology and Drug Safety*, 25(12):1414–1424.
- 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).
- Lisa Martin, Mike Hutchens, Craig Hawkins, et al. 2017. [How much do clinical trials cost?](#) *Nature Reviews Drug Discovery*, 16(6):381–382.
- Elizabeth McNeer, Cole Beck, Hannah L. Weeks, Michael L. Williams, Nathan T. James, and Leena Choi. 2019. [A post-processing algorithm for building longitudinal medication dose data from extracted medication information using natural language processing from electronic health records](#). *bioRxiv*.
- Rebecca M. Murphy, Joanna E. Klopotoska, Nicolette F. de Keizer, Kitty J. Jager, Jorg H. Leopold, Dave A. Dongelmans, et al. 2023.

- Adverse drug event detection using natural language processing: A scoping review of supervised learning methods. *PLoS ONE*, 18(1):e0279842.
- Senobar Naderian, Roya Rahmani, and Taha Samad-Soltani. 2024. A natural language processing framework for detecting adverse drug reactions in clinical structured drug reviews. *Int J Drug Res Clin*, 2(1):e9–e9.
- Ann Marie Navar, Michael J. Pencina, Jennifer A. Rymer, David M. Louzao, and Eric D. Peterson. 2016. Use of open access platforms for clinical trial data. *JAMA*, 315(12):1283–1284.
- Sai Nerella, Saptarshi Bandyopadhyay, Jiayu Zhang, Miguel Contreras, Samantha Siegel, Ahmet Bumin, Bruno Silva, Jose Sena, Benjamin Shickel, Azra Bihorac, Kamran Khezeli, and Parisa Rashidi. 2024. Transformers and large language models in healthcare: A review. *Artificial Intelligence in Medicine*, 154:102900.
- Nikhil Patwardhan, Stefano Marrone, and Carlo Sansone. 2023. Transformers in the real world: A survey on nlp applications. *Information*, 14(4):242.
- 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.
- J. G. Rim, J. G. Jackman, Christopher P. Hornik, Jonathan L. Rutter, Haider Warraich, Janet Wittes, Lee Fleisher, Brian S. Anderson, Edward Krofah, Eileen Kinter, Theresa Locke, Laura Kehoe, Aliya Abbasi, Hilary Marston, Ann Meeker-O’Connell, William Weber, T. Wang, Adrian F. Hernandez, Martin Landray, and Scott M. Palmer. 2024. Accelerating evidence generation: Addressing critical challenges and charting a path forward. *Journal of Clinical and Translational Science*, 8(1):e184.
- Janine P. Stewart. 2021. *Fda perspective on medication errors involving investigational drugs*. Reagan-Udall Foundation for the FDA.
- Rabia A. Tariq, Ritesh Vashisht, Ankur Sinha, and Yekaterina Scherbak. 2024. Medication dispensing errors and prevention. In *StatPearls*. StatPearls Publishing, Treasure Island (FL).
- U.S. National Library of Medicine. 2026. [Clinical-trials.gov](https://clinicaltrials.gov). Registry of clinical trials from around the world.
- Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12–20.
- Jetsun Whitton and Anthony Hunter. 2023. Automated tabulation of clinical trial results: A joint entity and relation extraction approach with transformer-based language representations. *Artificial Intelligence in Medicine*, 144:102661.
- Chi Heem Wong, Kien Wei Siah, and Andrew W. Lo. 2019. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286.
- Y. Zhu, L. Li, H. Lu, A. Zhou, and X. Qin. 2020. Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions. *Journal of Biomedical Informatics*, 106:103451.

## A. Appendices

### A.1. Full dataset column description

| Feature             | Feature meaning                            | Relevance to modelling                                                                                         |
|---------------------|--------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| phases              | Clinical trial phase (e.g. Phase 1–4)      | Indicates development stage, regulatory maturity, safety vs efficacy focus, and expected sample size.          |
| enrollmentCount     | Number of participants planned or enrolled | Proxy for trial scale, statistical power, and development stage; larger trials typically reflect later phases. |
| allocation          | Method of assigning participants to arms   | Core indicator of internal validity and risk of bias; randomisation underpins causal inference.                |
| interventionModel   | Structural design of the trial             | Determines treatment comparison logic and statistical assumptions.                                             |
| primaryPurpose      | Main intent of the trial                   | Distinguishes exploratory from confirmatory trials.                                                            |
| masking             | Level of blinding                          | Directly affects bias risk.                                                                                    |
| healthyVolunteers   | Recruitment of healthy participants        | Differentiates early-phase from therapeutic trials.                                                            |
| sex                 | Sex eligibility criteria                   | Important for generalisability.                                                                                |
| oversightHasDmc     | Presence of Data Monitoring Committee      | Indicates safety oversight rigor.                                                                              |
| briefSummary        | Plain-language study summary               | Useful for NLP-based screening.                                                                                |
| detailedDescription | Expanded study description                 | Rich source of procedural details.                                                                             |
| conditions          | Diseases studied                           | Enables disease-specific analysis.                                                                             |
| protocolPdfText     | Extracted protocol PDF text                | Contains detailed operational information.                                                                     |
| numArms             | Number of study arms                       | Reflects design complexity.                                                                                    |
| armDescriptions     | Free-text arm descriptions                 | Encodes dose and schedule information.                                                                         |
| interventionNames   | Names of interventions                     | Enables cross-trial drug analysis.                                                                             |
| numLocations        | Number of study sites                      | Proxy for geographical spread.                                                                                 |
| target              | Binary dosing error indicator              | Supervised learning target.                                                                                    |
| nctid               | Clinical trial identifier                  | Enables registry linkage.                                                                                      |

Table 1: Description of Dataset Features

### A.2. Character and token length of text columns

| Field                    | Count  | Mean   | Std      | Min | 25% | 50% | 75%   | Max    |
|--------------------------|--------|--------|----------|-----|-----|-----|-------|--------|
| briefSummary             | 29,478 | 491.09 | 459.66   | 16  | 209 | 330 | 592   | 4,900  |
| detailedDescription      | 29,478 | 967.18 | 1,505.99 | 0   | 0   | 457 | 1,369 | 30,703 |
| armDescriptions          | 29,478 | 347.02 | 496.59   | 0   | 78  | 197 | 425   | 13,350 |
| interventionDescriptions | 29,478 | 282.01 | 332.59   | 17  | 89  | 178 | 347   | 8,745  |

Table 2: Character Length Distribution of Textual Fields

| Field                    | Mean   | Median | Min | Max   | Std    | 25% | 50% | 75% | % > 256 |
|--------------------------|--------|--------|-----|-------|--------|-----|-----|-----|---------|
| briefSummary             | 117.39 | 81     | 5   | 1,231 | 105.12 | 52  | 81  | 142 | 8.65    |
| detailedDescription      | 228.85 | 111    | 2   | 6,917 | 350.95 | 2   | 111 | 327 | 31.81   |
| armDescriptions          | 100.12 | 59     | 2   | 4,400 | 138.71 | 27  | 59  | 122 | 7.79    |
| interventionDescriptions | 75.81  | 48     | 6   | 2,297 | 89.17  | 24  | 48  | 94  | 4.03    |

Table 3: Token Length Distribution of Textual Fields Using the BioBERT Tokenizer (Training subset)

### A.3. Statistical Analysis Categorical Features

| Input Feature     | $\chi^2$ Statistic | $p$ -value              | Cramér's V | Min Exp. Count | $\chi^2$ AND Cramér's V |
|-------------------|--------------------|-------------------------|------------|----------------|-------------------------|
| interventionModel | $2.19 \times 10^2$ | $2.28 \times 10^{-45}$  | 0.086      | 5.183          | TRUE                    |
| allocation        | $9.09 \times 10^1$ | $1.42 \times 10^{-19}$  | 0.056      | 6.926          | TRUE                    |
| oversightHasDmc   | $7.68 \times 10^1$ | $2.07 \times 10^{-17}$  | 0.051      | 176.854        | TRUE                    |
| armGroupTypes     | $1.19 \times 10^3$ | $2.53 \times 10^{-26}$  | 0.201      | 0.046          | FALSE                   |
| phases_new        | $9.60 \times 10^2$ | $5.50 \times 10^{-202}$ | 0.180      | 0.046          | FALSE                   |
| interventionTypes | $8.78 \times 10^2$ | $1.00 \times 10^0$      | 0.173      | 0.046          | FALSE                   |
| healthyVolunteers | $1.66 \times 10^2$ | $8.73 \times 10^{-37}$  | 0.075      | 1.238          | FALSE                   |
| primaryPurpose    | $1.25 \times 10^2$ | $1.16 \times 10^{-22}$  | 0.065      | 0.138          | FALSE                   |
| numArms           | 117.849            | $1.17 \times 10^{-13}$  | 0.063      | 0.046          | FALSE                   |
| masking           | 55.300             | $1.13 \times 10^{-10}$  | 0.043      | 3.211          | FALSE                   |
| sex               | 25.387             | $1.28 \times 10^{-5}$   | 0.029      | 0.321          | FALSE                   |

Table 4: Chi-square association analysis between categorical input features and

### A.4. GridSearch - Hyperparameter Search Space

| Model               | Hyperparameter    | Values               |
|---------------------|-------------------|----------------------|
| LightGBM            | num_leaves        | {31, 63}             |
|                     | max_depth         | {6, 8}               |
|                     | learning_rate     | {0.05, 0.1}          |
|                     | min_child_samples | {20, 50}             |
| Logistic Regression | C                 | {0.1, 1, 10}         |
|                     | penalty           | L2                   |
|                     | solver            | saga                 |
| SVC (RBF)           | C                 | {0.1, 1, 10}         |
|                     | gamma             | {scale, 0.01, 0.001} |
|                     | kernel            | RBF                  |

Table 5: Hyperparameter Search Spaces



## A.5. Best Models Stratified K-Fold Cross Validation Results

| Metric            | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Mean $\pm$ Std    |
|-------------------|--------|--------|--------|--------|--------|-------------------|
| ROC-AUC           | 0.728  | 0.722  | 0.740  | 0.737  | 0.733  | $0.732 \pm 0.006$ |
| F1 (macro)        | 0.570  | 0.557  | 0.567  | 0.568  | 0.563  | $0.565 \pm 0.005$ |
| Precision         | 0.150  | 0.138  | 0.149  | 0.149  | 0.144  | $0.146 \pm 0.004$ |
| Recall            | 0.627  | 0.633  | 0.662  | 0.652  | 0.651  | $0.645 \pm 0.013$ |
| Balanced Accuracy | 0.728  | 0.722  | 0.740  | 0.737  | 0.733  | $0.732 \pm 0.006$ |

Table 6: Stratified k-fold (k=5) Cross-Validation Results — Calibrated SVC

| Metric            | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Mean $\pm$ Std    |
|-------------------|--------|--------|--------|--------|--------|-------------------|
| ROC-AUC           | 0.842  | 0.833  | 0.853  | 0.837  | 0.843  | $0.841 \pm 0.007$ |
| F1 (macro)        | 0.514  | 0.517  | 0.528  | 0.514  | 0.522  | $0.519 \pm 0.005$ |
| Precision         | 0.429  | 0.526  | 0.500  | 0.375  | 0.429  | $0.452 \pm 0.054$ |
| Recall            | 0.028  | 0.031  | 0.043  | 0.027  | 0.037  | $0.033 \pm 0.006$ |
| Balanced Accuracy | 0.513  | 0.515  | 0.520  | 0.513  | 0.517  | $0.516 \pm 0.003$ |

Table 7: Stratified k-fold (k=5) Cross-Validation Results — Logistic Regression

| Metric            | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Mean $\pm$ Std    |
|-------------------|--------|--------|--------|--------|--------|-------------------|
| ROC-AUC           | 0.833  | 0.825  | 0.849  | 0.826  | 0.845  | $0.836 \pm 0.010$ |
| F1 (macro)        | 0.599  | 0.589  | 0.594  | 0.589  | 0.602  | $0.595 \pm 0.005$ |
| Precision         | 0.181  | 0.169  | 0.175  | 0.169  | 0.183  | $0.175 \pm 0.006$ |
| Recall            | 0.526  | 0.523  | 0.567  | 0.524  | 0.596  | $0.547 \pm 0.030$ |
| Balanced Accuracy | 0.706  | 0.700  | 0.719  | 0.700  | 0.735  | $0.712 \pm 0.013$ |

Table 8: Stratified k-fold (k=5) Cross-Validation Results — LightGBM

| Metric            | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Mean $\pm$ Std    |
|-------------------|--------|--------|--------|--------|--------|-------------------|
| ROC-AUC           | 0.816  | 0.825  | 0.853  | 0.827  | 0.834  | $0.831 \pm 0.012$ |
| F1 (macro)        | 0.564  | 0.568  | 0.568  | 0.594  | 0.571  | $0.573 \pm 0.011$ |
| Precision         | 0.145  | 0.146  | 0.151  | 0.175  | 0.150  | $0.153 \pm 0.011$ |
| Recall            | 0.639  | 0.563  | 0.701  | 0.549  | 0.612  | $0.613 \pm 0.055$ |
| Balanced Accuracy | 0.729  | 0.702  | 0.756  | 0.712  | 0.723  | $0.725 \pm 0.018$ |

Table 9: Stratified k-fold (k=5) Cross-Validation Results — Text-Tower Neural Network

## A.6. Neural network architectures

### A.6.1. Simple Multi-layer Perceptron:

```
def build_basic_binary_classifier(input_dim):
 model = keras.Sequential([
 layers.Input(shape=(input_dim,)),
 layers.Dense(128, activation="relu"),
 layers.Dropout(0.3),
 layers.Dense(64, activation="relu"),
 layers.Dropout(0.2),
 layers.Dense(1, activation="sigmoid")
])
 return model
```

### A.6.2. Wide Multi-layer Perceptron:

```
def build_binary_classifier_wide(input_dim):
 model = keras.Sequential([
 layers.Input(shape=(input_dim,)),

 # Wider first layer to capture feature interactions
 layers.Dense(512, activation="relu", kernel_regularizer=keras.regularizers.l2(0.001)),
 layers.BatchNormalization(),
 layers.Dropout(0.4),

 layers.Dense(256, activation="relu", kernel_regularizer=keras.regularizers.l2(0.001)),
 layers.BatchNormalization(),
 layers.Dropout(0.3),

 layers.Dense(128, activation="relu", kernel_regularizer=keras.regularizers.l2(0.001)),
 layers.Dropout(0.2),

 layers.Dense(64, activation="relu"),
 layers.Dropout(0.2),

 layers.Dense(1, activation="sigmoid")
])
 return model
```

### A.6.3. Residual connections/skip layer:

```
def build_binary_classifier_skip_layers(input_dim):
 inputs = layers.Input(shape=(input_dim,))

 # First block
 x = layers.Dense(512, activation="relu", kernel_regularizer=keras.regularizers.l2(0.001))(inputs)
 x = layers.BatchNormalization()(x)
 x = layers.Dropout(0.4)(x)

 # Second block with skip connection
 x2 = layers.Dense(256, activation="relu", kernel_regularizer=keras.regularizers.l2(0.001))(x)
 x2 = layers.BatchNormalization()(x2)
 x2 = layers.Dropout(0.3)(x2)

 # Third block
 x3 = layers.Dense(128, activation="relu", kernel_regularizer=keras.regularizers.l2(0.001))(x2)
```

```

x3 = layers.BatchNormalization()(x3)
x3 = layers.Dropout(0.2)(x3)

Concatenate different levels (multi-scale features)
x2_pooled = layers.Dense(64, activation="relu")(x2)
x3_pooled = layers.Dense(64, activation="relu")(x3)
combined = layers.Concatenate()([x2_pooled, x3_pooled])

Final layers
x = layers.Dense(64, activation="relu")(combined)
x = layers.Dropout(0.2)(x)
outputs = layers.Dense(1, activation="sigmoid")(x)

model = keras.Model(inputs=inputs, outputs=outputs)
return model

```

#### A.6.4. Text tower model:

```

def build_single_input_text_cat_fusion_model(
 input_dim: int,
 text_embedding_dim: int = 768,
 num_text_fields: int = 4,
 num_cat_features: int | None = None, # if None, inferred as input_dim -
 num_text_fields * text_embedding_dim
 text_proj_dim: int = 192,
 cat_proj_dim: int = 64,
 hidden_dim: int = 128,
 dropout_rate: float = 0.2,
 l2_strength: float = 1e-4,
 use_gating: bool = True,
):
 """
 Single-input model: expects input vector shaped (batch, input_dim)
 Layout: [text1(768)|text2(768)|text3(768)|text4(768)|cat_features (...)]
 """
 l2 = keras.regularizers.L2(l2_strength)

 if num_cat_features is None:
 num_cat_features = input_dim - (num_text_fields * text_embedding_dim)
 if num_cat_features < 0:
 raise ValueError(
 f"input_dim={input_dim} is smaller than {(num_text_fields)*text_embedding_dim}="
 f"{num_text_fields*text_embedding_dim}"
)

 inputs = layers.Input(shape=(input_dim,), name="flat_input")

 # ---- Slice flat vector into blocks ----
 def slice_block(start_idx, end_idx, name):
 return layers.Lambda(lambda x: x[:, start_idx:end_idx], name=name)(inputs)

 text_blocks = []
 for i in range(num_text_fields):
 start = i * text_embedding_dim
 end = start + text_embedding_dim
 text_blocks.append(slice_block(start, end, name=f"text_{i+1}_slice"))

 cat_start = num_text_fields * text_embedding_dim

```

```

cat_end = cat_start + num_cat_features
cat_block = slice_block(cat_start, cat_end, name="cat_slice")

---- Per-text-field towers ----
text_reprs = []
for i, t in enumerate(text_blocks):
 x = layers.LayerNormalization(name=f"text_{i+1}_ln")(t)
 x = layers.Dense(text_proj_dim, activation="gelu", kernel_regularizer=l2,
 name=f"text_{i+1}_proj1")(x)
 x = layers.Dropout(dropout_rate, name=f"text_{i+1}_drop1")(x)
 x = layers.Dense(text_proj_dim, activation="gelu", kernel_regularizer=l2,
 name=f"text_{i+1}_proj2")(x)
 text_reprs.append(x) # (batch, text_proj_dim)

---- Optional gating over text fields ----
if use_gating:
 stacked = layers.Concatenate(name="text_concat_for_gate")(text_reprs) # (batch,
 num_text_fields*text_proj_dim)
 gate = layers.Dense(64, activation="relu", kernel_regularizer=l2,
 name="gate_dense")(stacked)
 gate = layers.Dense(num_text_fields, activation="softmax", name="gate_weights")(gate) #
 (batch, 4)

 weighted = []
 for i in range(num_text_fields):
 w_i = layers.Lambda(lambda g, idx=i: tf.expand_dims(g[:, idx], axis=-1),
 name=f"gate_w_{i+1}")(gate)
 weighted.append(layers.Multiply(name=f"text_{i+1}_weighted")([text_reprs[i], w_i]))

 text_fused = layers.Add(name="text_fused")(weighted) # (batch, text_proj_dim)
else:
 # simpler: just concatenate all text representations
 text_fused = layers.Concatenate(name="text_fused_concat")(text_reprs)

---- Categorical tower ----
c = layers.LayerNormalization(name="cat_ln")(cat_block)
c = layers.Dense(cat_proj_dim, activation="relu", kernel_regularizer=l2, name="cat_proj")(c)
c = layers.Dropout(dropout_rate, name="cat_drop")(c)

---- Fusion head ----
x = layers.Concatenate(name="fusion_concat")([text_fused, c])
x = layers.Dense(hidden_dim, activation="relu", kernel_regularizer=l2,
 name="fusion_dense1")(x)
x = layers.Dropout(dropout_rate, name="fusion_drop1")(x)
x = layers.Dense(hidden_dim // 2, activation="relu", kernel_regularizer=l2,
 name="fusion_dense2")(x)
x = layers.Dropout(dropout_rate, name="fusion_drop2")(x)

outputs = layers.Dense(1, activation="sigmoid", name="output")(x)

model = keras.Model(inputs=inputs, outputs=outputs, name="single_input_text_cat_fusion")
return model

```

# CGU-ILALab at FoodBench-QA 2026: Comparing Traditional and LLM-based Approaches for Recipe Nutrient Estimation

Wei-Chun Chen<sup>a</sup>, Yu-Xuan Chen<sup>b</sup>, I-Fang Chung<sup>a</sup>, Ying-Jia Lin<sup>c,d</sup>

<sup>a</sup>Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, Taiwan

<sup>b</sup>Institute of Health Data Science, Chang Gung University, Taoyuan, Taiwan

<sup>c</sup>Department of Artificial Intelligence, Chang Gung University, Taoyuan, Taiwan

<sup>d</sup>Artificial Intelligence Research Center, Chang Gung University, Taoyuan, Taiwan

{andybox11111.md14, ifchung}@nycu.edu.tw, {m1462004, yjlin}@cgu.edu.tw

## Abstract

Accurate nutrient estimation from unstructured recipe text is an important yet challenging problem in dietary monitoring, due to ambiguous ingredient terminology and highly variable quantity expressions. We systematically evaluate models spanning a wide range of representational capacity, from lexical matching methods (TF-IDF with Ridge Regression), to deep semantic encoders (DeBERTa-v3), to generative reasoning with large language models (LLMs). Under the strict tolerance criteria defined by EU Regulation 1169/2011, our empirical results reveal a clear trade-off between predictive accuracy and computational efficiency. The TF-IDF baseline achieves moderate nutrient estimation performance with near-instantaneous inference, whereas the DeBERTa-v3 encoder performs poorly under task-specific data scarcity. In contrast, few-shot LLM inference (e.g., Gemini 2.5 Flash) and a hybrid LLM refinement pipeline (TF-IDF combined with Gemini 2.5 Flash) deliver the highest validation accuracy across all nutrient categories. These improvements likely arise from the ability of LLMs to leverage pre-trained world knowledge to resolve ambiguous terminology and normalize non-standard units, which remain difficult for purely lexical approaches. However, these gains come at the cost of substantially higher inference latency, highlighting a practical deployment trade-off between real-time efficiency and nutritional precision in dietary monitoring systems.

**Keywords:** Recipe Nutrient Estimation, Large Language Models

## 1. Introduction

Accurate nutrient estimation from recipe text is a practically relevant yet technically challenging task in the food and nutrition domain. While large language models (LLMs) have demonstrated remarkable capabilities in general-purpose reasoning and knowledge retrieval, it remains unclear how well they handle the structured, quantitative reasoning required for food applications. Recipe texts present unique difficulties: ambiguous ingredient terminology, non-standard measurement units, and the need to map diverse culinary expressions to precise nutritional values. These challenges make recipe nutrient estimation a compelling testbed for evaluating whether different modeling paradigms — from traditional lexical methods to LLM-based inference — can produce reliable nutritional predictions.

The FoodBench-QA 2026 shared task<sup>1</sup> asks systems to answer food and nutrition questions based on unstructured recipe text. The complexity of the food domain stems from its inherent heterogeneity. Unlike standardized datasets in other fields, culinary data is characterized by ambiguous ingredient terminology, non-standard and highly variable measurement units (e.g., “a pinch,” “a medium bunch”), and the challenge of mapping these varied expres-

sions to structured food composition databases. A model must not only perform accurate entity disambiguation to distinguish between semantically similar but nutritionally distinct ingredients (e.g., “coconut milk” vs. “coconut water”) but also execute complex reasoning to normalize quantities and aggregate nutritional values from disparate sources.

---

**Case 1 Prompt:** Assess the nutritional profile of 180 ml of pineapple juice, canned or bottled, unsweetened, without added ascorbic acid.

**Case 1 Answer:** The nutrient values demonstrated here are: energy - 99.79, fat - 0.23, protein - 0.68, salt - 3.78, saturates - 0.02, sugars - 18.86.

---

**Case 2 Prompt:** Check the nutritional values per 100 g in a recipe that comprises these ingredients: 2 teaspoons corn, sweet, white, raw, 2 teaspoons corn, sweet, white, raw, 2 teaspoons peppers, sweet, green, raw, 1 teaspoon spices, coriander seed, 1/2 cup butter, without salt.

**Case 2 Answer:** Nutrient details in 100 g: energy - 713.49, fat - 80.53, protein - 0.86, salt - 0.04, saturates - 50.10, sugars - 0.09.

---

Table 1: Examples in the FoodBench-QA 2026 shared task. Prompt: nutritional queries; Answer: response that should be provided by a system.

In response to these challenges, the FoodBench-

---

<sup>1</sup><https://www.codabench.org/competitions/12112/>



QA 2026 shared task calls for systems capable of moving beyond simple pattern matching to provide accurate nutritional estimates that comply with regulatory standards. Motivated by this critical need, this paper focuses on the sub-task of nutrient estimation for recipes, where examples from this sub-task are shown in Table 1. Though each data label includes six target nutrients (Energy, Salt, Fat, Protein, Saturates, Sugars), the FoodBench-QA shared task focused on evaluation of the latter four nutrients. To tackle the “real-world” noise of dietary data, we systematically explore and evaluate multiple methodologies. By investigating various approaches to resolving ambiguity and standardizing measurements, we aim to identify the most effective strategies<sup>2</sup> for improving nutrient estimation accuracy, thereby advancing the reliability of automated systems in health-oriented language technologies.

## 2. Dataset

The dataset utilized in this study is sourced from the “FoodBench-QA: Shared Task on Grounded Food & Nutrition Question Answering” competition on Codabench. Specifically, we employed the annotated subset of the data without titles (**T1.1 - Ingredients**)<sup>3</sup>. During the preprocessing stage, we identified and removed duplicate recipe entries to ensure data quality. After the data cleaning process, a total of 14,512 unique samples remained. Finally, we partitioned the dataset into our training and validation sets at an 80:20 ratio, resulting in 11,609 and 2,903 examples for training and validation, respectively.

## 3. Method

To comprehensively investigate the applicability of models with varying complexities to this task, this study systematically compares multiple approaches ranging from traditional TF-IDF pipelines to LLM-based inference, as well as a hybrid strategy combining TF-IDF with an LLM.

Starting with fundamental lexical statistical methods, we implemented an approach combining TF-IDF (Salton and Buckley, 1988) with traditional machine learning. Subsequently, we advanced into the deep learning domain by employing an encoder-based model (He et al., 2023) to enhance the capability for semantic understanding. Finally, we extended our research to current state-of-the-art

technologies by directly utilizing large language models (LLMs) for few-shot inference. This design not only establishes a clear comparative baseline but also explicitly demonstrates the impact of technological evolution on overall task performance.

### 3.1. Term Frequency-Inverse Document Frequency (TF-IDF) with Machine Learning

As our foundational baseline, we represent recipe texts as sparse lexical vectors using TF-IDF (Salton and Buckley, 1988), coupled with Ridge Regression (Hoerl and Kennard, 1970) for the continuous prediction tasks. TF-IDF is a widely adopted representation for converting text into numerical vectors in text-based prediction tasks (Sparck Jones, 1972).

TF-IDF features are constructed using recipe ingredient text as input, with basic preprocessing including lowercasing and the removal of English stop words at the word level. To capture both broader semantic information and subtle spelling variations inherent in culinary data, we extract two distinct types of TF-IDF features.

First, **Word-level TF-IDF** applies unigram and bigram features to capture key ingredient terms and common word combinations. To control feature quality and reduce noise, a minimum document frequency of  $\text{min\_df} = 2$  is applied, retaining only terms that appear in at least two documents. A maximum document frequency of  $\text{max\_df} = 0.9$  removes overly common terms appearing in more than 90% of documents. The vocabulary size is limited to 8,000 features, and logarithmic scaling is applied to term frequencies to dampen the influence of high-frequency terms.

Second, following the character n-gram approach proposed by Cavnar et al. (1994), **Character-level TF-IDF** utilizes 3–5 word-boundary-aware character n-grams to effectively handle spelling variations, typos, and fine-grained morphological patterns. For these features,  $\text{min\_df} = 2$  and  $\text{max\_df} = 0.95$  are applied, and the feature size is limited to 12,000 to balance expressiveness and dimensionality.

By combining these word-level and character-level features, we form a comprehensive final feature vector. This hybrid representation retains both word-level meaning and character-level details, making the model highly robust to variations in textual expression.

Furthermore, Ridge Regression, with its  $L_2$  regularization, is particularly well-suited for handling the resulting high-dimensional and sparse feature space. Critically, it mitigates the multicollinearity naturally introduced by overlapping word and character features. By penalizing large coefficients,

<sup>2</sup>Our code: <https://github.com/ila-lab/FoodBench-QA>

<sup>3</sup><https://github.com/matejMartinc/FoodBench-QA-train/tree/main/recipe%20nutrient%20assessment/ingredients>

Ridge Regression prevents the model from over-concentrating its “attention” (i.e., assigning extreme weights) solely on perfectly matched words. Instead, it encourages a more balanced weight distribution, ensuring the model genuinely leverages character-level features to handle textual noise without overfitting.

Consequently, this methodology not only provides computational efficiency but also offers high interpretability, making it a strong baseline for ingredient terminology that can be matched explicitly.

### 3.2. Encoder-based NLP Models

To overcome the limitations of sparse representations, we advance to deep context-aware architectures using pre-trained bidirectional encoders. Specifically, we implement **DeBERTa-v3** (Decoding-enhanced BERT with disentangled attention) (He et al., 2023) as our core model for this stage. DeBERTa-v3 significantly improves upon standard BERT architectures (Devlin et al., 2019) by utilizing a disentangled attention mechanism, which encodes words and their relative positions separately. In principle, this architecture could benefit culinary data, where the relative positioning of numbers, units, and ingredients (e.g., “2 tablespoons of olive oil” vs. “oil, 2 tablespoons”) carries important semantic information.

We directly feed the raw ingredient text into the model without additional preprocessing or manual extraction. The model processes this sequence to generate context-aware embeddings before passing them to a linear regression head to predict nutritional values.

In theory, this deep learning tier could address the challenge of ambiguous ingredient terminology highlighted earlier. By dynamically capturing the surrounding text and positional context, DeBERTa-v3 may facilitate implicit word sense disambiguation, potentially differentiating between semantically similar but nutritionally disparate items (e.g., inferring whether “coconut” refers to milk, water, or oil based on the recipe context). However, as we discuss in the results, these theoretical advantages did not materialize in our low-data setting.

### 3.3. Large Language Models (LLMs): Direct Inference and Semantic Refinement

Finally, we extend our framework to the current state-of-the-art by utilizing generative LLMs. In this tier, we explore the capabilities of LLMs through two distinct operational paradigms: direct inference and post-processing refinement.

For **Direct Inference**, we evaluate a suite of locally deployed open-source LLMs, specifically

GPT-OSS-20B (OpenAI, 2025)<sup>4</sup>, Gemma-3-27B (Gemma Team, 2025)<sup>5</sup>, and Nemotron-3-Nano-30B (NVIDIA, 2025)<sup>6</sup>, via few-shot prompting. In addition, we evaluate Gemini 2.5 Flash (Comanici et al., 2025) as a cloud-based direct inference model using the same few-shot prompting strategy, to compare API-served models against locally deployed ones. The prompt design for direct inference, which includes two in-context examples (i.e., 2-shot), is shown in Figure 1. Instead of traditional fine-tuning, we task these models with implicitly executing the entire estimation pipeline, encompassing ingredient extraction, unit normalization, and nutrient aggregation, within a single inference step. These models are particularly well-suited to address the persistent issue of non-standard and highly variable measurement units (e.g., rationally transforming “a medium bunch” or “a pinch” into quantifiable metrics). By leveraging their pre-trained world knowledge, they perform the multi-step reasoning required to bridge the gap between noisy culinary instructions and nutritional estimations.

Complementing this direct approach, we introduce an **LLM-based Refinement** strategy to compensate for the limited semantic understanding of our shallower baseline models (e.g., the TF-IDF approach). For this post-processing step, we utilize Gemini 2.5 Flash (Comanici et al., 2025), with the prompt shown in Figure 2. For each test sample, the original ingredient text alongside the initial numerical predictions generated by the baseline model are provided as context to the LLM. The model evaluates whether the predicted values fall within reasonable nutritional boundaries and adjusts them when necessary. This step effectively acts as a semantic refinement layer, helping to align purely statistical or lexical predictions with real-world culinary logic and nutritional common sense.

## 4. Results

In this section, we present the empirical results comparing multiple modeling approaches for recipe nutrient estimation. We evaluate all approaches using the official evaluation tool, which operationalizes EU Regulation 1169/2011 (European Parliament and Council, 2011) tolerance thresholds as binary accuracy criteria — a more practically meaningful standard than continuous error metrics such as MAE or RMSE, as it directly reflects whether

<sup>4</sup><https://huggingface.co/openai/gpt-oss-20b>

<sup>5</sup><https://huggingface.co/google/gemma-3-27b-it>

<sup>6</sup><https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Nano-30B-A3B-BF16>

**System Role:**

You are a specialized nutritional data analyst. Your task is to calculate the nutrient profile per 100g for recipes provided in [INST] format.

**Instructions:**

**Unit Conversion:** Convert all units (e.g., pounds, cups, tablespoons, ml) to grams (g) using standard conversion factors (e.g., 1 cup water  $\approx$  236.6g, 1 tablespoon butter  $\approx$  14.2g).

**Calculation:** Sum the total weight and total nutrients of all ingredients, then normalize the values to a 100g portion.

**Output Format:** You must only provide the final result in this specific format:

```
Nutrient values per 100 g: fat -
[value], protein - [value], satu-
rates - [value], sugars - [value]
```

**input1:** [INST] Identify the nutritional content per 100 grams for a recipe with the following ingredients: 2 tablespoon soy sauce made from soy (tamari), 1 tablespoon peanut butter, smooth style, without salt, ... [ingredients truncated for brevity, paper display only] [/INST]

**answer1:** Nutrient values per 100 g: fat - 8.55, protein - 12.31, saturates - 1.72, sugars - 14.17

**input2:** [INST] Identify the nutritional content per 100 grams for a recipe with the following ingredients: 1 cup wheat flour, 2 tbsp olive oil, ... [ingredients truncated for brevity, paper display only] [/INST]

**answer2:** Nutrient values per 100 g: fat - 14.20, protein - 3.10, saturates - 2.15, sugars - 0.50

Figure 1: The prompt design used for direct inference with an LLM.

a model's predictions fall within the margins that regulators deem acceptable for food labeling.

## 4.1. Performance Comparison

As presented in Table 2, our results reveal a consistent pattern: models with greater pre-trained world knowledge achieve higher accuracy, while lightweight approaches retain a decisive advantage in inference speed.

We first evaluated the TF-IDF and Ridge Regression pipeline as our baseline. This approach achieved notable performance on Protein (64.62), while Fat (40.58), Saturates (50.33), and Sugar (50.26) remained at a moderate level. These results suggest that explicit lexical matching provides a reliable signal for nutrient estimation even without any semantic understanding.

To capture deeper semantic relationships, we subsequently evaluated the DeBERTa-v3 encoder. However, contrary to the typical dominance of deep encoder-based models in NLP tasks, DeBERTa-v3 performed substantially worse than the other

You are a nutrition expert.

**Food:**

{text}

**Predicted nutrients per 100g:**

Protein: {pred['protein\_g']:.2f}

Fat: {pred['fat\_g']:.2f}

Sugar: {pred['sugars\_g']:.2f}

Saturates: {pred['saturates\_g']:.2f}

**Return JSON only with keys:**

protein\_g, fat\_g, sugars\_g, saturates\_g

Figure 2: The prompt template used to refine nutrient predictions. Variables in monospaced font are dynamically filled with an input query ({text}) and TF-IDF prediction values during inference.

approaches. This underperformance likely stems from the limited scale of task-specific training data relative to DeBERTa-v3's parameter count, leaving the model insufficient signal to fine-tune its representations effectively for this regression task. Deep learning architectures typically require substantially larger datasets to learn effective representations for domain-specific tasks, and our constrained culinary dataset may not provide enough supervision for a model of this capacity.

To bypass the limitations of sparse training data, we shifted our paradigm to large language models (LLMs) utilizing few-shot inference. This approach yielded substantial improvements. Among the open-source models deployed locally, Gemma-3-27B achieved the strongest results, with scores of 60.21 for Protein and 48.02 for Saturates. Gemini 2.5 Flash, accessed via API, further surpassed all local models, achieving the highest scores for Fat (54.12) and Saturates (59.73). These results suggest that LLMs can better handle complex culinary reasoning, possibly by leveraging broad pre-trained knowledge for unit normalization and ingredient disambiguation.

Given that the TF-IDF baseline showed reasonable performance despite its simplicity, we hypothesized that augmenting it with LLM reasoning could offer further gains. As shown in the table, employing Gemini 2.5 Flash as a semantic refinement layer on top of the TF-IDF outputs successfully boosted the baseline's performance across all metrics (e.g., improving Protein from 64.62 to 67.79). This supports the viability of a hybrid approach where TF-IDF captures lexical co-occurrence patterns from seen ingredients, while the LLM corrects predictions where semantic understanding (such as unit normalization or ingredient disambiguation) is required.

| Model Architecture           | Performance Metrics |         |       |           | Latency |
|------------------------------|---------------------|---------|-------|-----------|---------|
|                              | Sugar               | Protein | Fat   | Saturates |         |
| Traditional Machine Learning |                     |         |       |           |         |
| TF-IDF + Ridge Regression    | 50.26               | 64.62   | 40.58 | 50.33     | 1 ms    |
| Encoder-based model          |                     |         |       |           |         |
| DeBERTa-v3                   | 8.87                | 23.86   | 8.27  | 8.01      | 3.58 ms |
| LLM-based Direct Inference   |                     |         |       |           |         |
| GPT-OSS-20B                  | 32.07               | 41.38   | 29.93 | 37.00     | 9.9 s   |
| Gemma-3-27B                  | 41.51               | 60.21   | 41.44 | 48.02     | 1.4 s   |
| Nemotron-3-Nano-30B          | 22.05               | 30.58   | 19.73 | 26.23     | 23.7 s  |
| Gemini 2.5 Flash             | 55.36               | 67.17   | 54.12 | 59.73     | 1.0 s   |
| LLM-based Refinement         |                     |         |       |           |         |
| TF-IDF + Gemini 2.5 Flash    | 55.46               | 67.79   | 52.12 | 58.39     | 1.0 s   |

Table 2: Performance in accuracy (%) and computational cost of evaluated methods. Metrics follow EU Regulation 1169/2011 tolerance criteria. Bold indicates best performance per metric.

| Model Architecture                | Performance Metrics |              |              |              |
|-----------------------------------|---------------------|--------------|--------------|--------------|
|                                   | Sugar               | Protein      | Fat          | Saturates    |
| <i>Encoder-based model</i>        |                     |              |              |              |
| DeBERTa-v3                        | 8.08                | 28.25        | 9.05         | 8.44         |
| <i>LLM-based Direct Inference</i> |                     |              |              |              |
| Gemma-3-27B                       | 42.43               | 59.17        | 38.35        | 46.41        |
| Gemma-3-27B + GPT-OSS-20B         | <b>43.03</b>        | <b>59.71</b> | <b>39.21</b> | <b>47.13</b> |

Table 3: Performance comparison of DeBERTa-v3 and the LLMs on the final test set. Metrics follow EU Regulation 1169/2011 tolerance criteria. Bold indicates best performance per metric.

## 4.2. Computational Efficiency and Trade-offs

To ensure a consistent and standardized evaluation of computational efficiency, all local model inferences—spanning the TF-IDF pipeline, DeBERTa-v3, and the open-source LLMs (GPT-OSS-20B, Gemma-3-27B, and Nemotron-3-Nano-30B)—were executed on a dedicated workstation equipped with a single **NVIDIA RTX PRO 4500 (Blackwell architecture) GPU**. For the hybrid TF-IDF + Gemini 2.5 Flash pipeline, latency measurements encompass the API network routing and server-side generation time.

While predictive accuracy is crucial, practical deployment in dietary monitoring necessitates a strict evaluation of this inference latency. The rightmost column of Table 2 highlights a stark computational trade-off.

The TF-IDF baseline is exceptionally lightweight, boasting a near-instantaneous inference time of 1 ms per recipe. While DeBERTa-v3 is also highly

efficient (3.58 ms), its poor predictive performance disqualifies it for this specific low-data task.

Conversely, the superior predictive accuracy of LLMs comes at a significant computational cost, even when accelerated by the Blackwell architecture. The few-shot inference times for the open-source LLMs range from 1.4 seconds (Gemma-3-27B) to 23.7 seconds (Nemotron-3-Nano-30B). Notably, Gemma-3-27B (27B parameters) achieves faster inference than the smaller GPT-OSS-20B (20B parameters) despite its larger model size. This is attributable to differences in quantization: Gemma-3-27B is deployed with Q4\_K\_M quantization, which enables efficient execution on consumer-grade hardware, whereas GPT-OSS-20B uses MXFP4 quantization, which incurs higher overhead on our hardware configuration. Similarly, both the Gemini 2.5 Flash direct inference and the TF-IDF + Gemini 2.5 Flash pipeline require approximately 1.0 second per request due to API overhead.

Ultimately, for latency-critical applications, the



pure TF-IDF model remains the most practical solution. For accuracy-prioritized scenarios, the Gemini 2.5 Flash-based approaches — both direct inference and the hybrid TF-IDF + Gemini pipeline — deliver the highest performance across all nutrient categories while maintaining a competitive latency of 1.0 second per request. When only local GPU resources are available without API access, Gemma-3-27B offers the best trade-off among open-source models at a moderate inference cost of 1.4 seconds.

## 5. Final Scores on the Test Set

Based on the results in Table 2, we submitted our predictions (Team Name: **andybox111**) for the final test set using DeBERTa-v3, Gemma-3-27B, and the hybrid predictions from Gemma-3-27B and GPT-OSS-20B. The hybrid prediction is based on 700 samples from GPT-OSS-20B and the rest from Gemma-3-27B.

Although the TF-IDF baseline achieved competitive validation scores, we prioritized neural and LLM-based systems for official submission. Similarly, the Gemini 2.5 Flash experiments — both direct inference and the hybrid TF-IDF + Gemini pipeline — were conducted after the submission deadline as part of our extended analysis for this paper. Furthermore, as a proprietary cloud API, Gemini 2.5 Flash’s outputs may vary across API versions, posing challenges for reproducibility.

The official evaluation results are presented in Table 3. Consistent with our validation findings, the LLM-based approaches significantly outperform the encoder-based model. The hybrid approach (Gemma-3-27B + GPT-OSS-20B) achieved the highest scores across all nutrient categories. However, its performance is only marginally better than the standalone Gemma-3-27B, likely due to the small number of predictions generated by GPT-OSS-20B.

## 6. Conclusion

In this study, we systematically compared three tiers of approaches — from lexical matching to encoder-based models to generative LLMs — for automated recipe nutrient estimation. Our experiments revealed that model complexity alone does not guarantee better performance; rather, the availability of pre-trained world knowledge appeared to be a key factor in this low-resource, domain-specific task.

First, we showed that a well-engineered TF-IDF baseline combined with Ridge Regression can achieve reasonable predictive performance with negligible computational overhead, making it a

practical choice for resource-constrained or real-time applications. Second, our experiments highlighted a critical limitation of standard deep learning architectures in specialized domains: models like DeBERTa-v3 may struggle under task-specific data scarcity, leading to severe performance degradation.

Finally, LLM-based approaches delivered the highest accuracy in our experiments. On the official test set, the open-source Gemma-3-27B and its hybrid with GPT-OSS-20B achieved the best submitted scores, suggesting that pre-trained world knowledge may help with complex culinary reasoning without task-specific fine-tuning. In our extended validation analysis, Gemini 2.5 Flash-based approaches further improved performance (e.g., the hybrid TF-IDF + Gemini pipeline improved Fat from 40.58 to 52.12 over the TF-IDF baseline), though these results should be interpreted with caution as they were not evaluated on the official test set and rely on a proprietary API whose outputs may not be fully reproducible.

However, the superior accuracy of LLMs introduces a tangible trade-off with computational efficiency, presenting distinct deployment choices depending on the application’s latency tolerance. Future work will explore model distillation and quantization techniques aiming to transfer the complex reasoning capabilities of LLMs into smaller, faster architectures, thereby achieving an optimal balance of accuracy and real-time efficiency for dietary monitoring systems.

## 7. Acknowledgements

We sincerely thank the reviewers for their valuable comments and constructive suggestions, which helped improve the quality of this work. This work was partially supported by the National Science and Technology Council, Taiwan, under Grant No. NSTC 114-2222-E-182-001-MY2.

## 8. Bibliographical References

- William B Cavnar, John M Trenkle, et al. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175, page 14. Las Vegas, NV.
- Gheorghe Comanici et al. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multi-modality, long context, and next generation agentic capabilities](#).
- 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.

European Parliament and Council. 2011. Regulation (eu) no 1169/2011 on the provision of food information to consumers. *Official Journal of the European Union*.

Gemma Team. 2025. [Gemma 3 technical report](#).

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*.

Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

NVIDIA. 2025. [Nemotron 3 nano: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning](#).

OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#).

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

Karen Sparck Jones. 1972. [A statistical interpretation of term specificity and its application in retrieval](#). *Journal of Documentation*, 28(1):11–21.

# FCProfiler: Structured and Deterministic Pipeline for Recipe-Level Nutrient Estimation

Siyeon Kim

sykim2601@knou.ac.kr

## Abstract

Food and nutrition question answering involves resolving ambiguous ingredient terminology and diverse household measurement expressions, and converting them into representations compatible with nutrient databases. In FoodBench-QA, recipe-level nutrient estimation requires consistent handling of heterogeneous and imprecise measurement descriptions. We propose FoodComponentProfiler (FCProfiler), a deterministic pipeline that treats nutrient estimation as a structured measurement resolution problem. The pipeline is composed of multiple stages, including parsing, normalization, unit canonicalization, gram conversion, and nutrient estimation, with each step designed to remain transparent and traceable. Unit canonicalization combines rule-based standards with data-driven unit expansion from large-scale recipe corpora, enabling broader coverage of real-world measurement variations. Gram conversion grounds quantities in ingredient-specific portion information, enabling accurate and traceable mass computation. Experimental results show that accurate nutrient estimation mainly depends on reliable unit normalization and ingredient-specific measurement conversion. Additionally, FCProfiler achieves performance comparable to FoodyLLM, demonstrating that explicit measurement grounding serves as an effective alternative to implicit reasoning. The proposed methodology preserves interpretability while maintaining strong performance in food and nutrition question answering.

**Keywords:** recipe-level nutrient estimation, measurement normalization, deterministic modeling

## 1. Introduction

Food and nutrition question answering involves challenges beyond general-purpose QA. It requires resolving ambiguous ingredient terminology, heterogeneous measurement expressions, and aligning textual descriptions with structured food composition databases (Min et al., 2019).

FoodBench-QA shared task, held as part of CL4Health at LREC 2026, addresses these challenges using nutrient databases and food ontologies. In FoodBench-QA, the problem is defined as a grounded food understanding task, where recipe-level nutrient estimation is evaluated under EU Regulation 1169/2011 tolerance standards (Bairati, 2017). This setting places strong emphasis on accurate mass normalization and database-grounded nutrient estimation.

In this context, consistent handling of measurement expressions becomes critical. Household units such as “cup,” “tablespoon,” or “pinch” are ingredient-dependent and loosely defined, and even small inconsistencies in unit interpretation or gram conversion can lead to substantial deviations in nutrient estimates, particularly for high-density nutrients.

While large language models (LLMs) demonstrate strong generative performance, end-to-end generation provides limited control over numerical normalization and intermediate mass computation, which are essential for reproducible nutrient estimation (Minaee et al., 2025).

To address these challenges, we develop FoodComponentProfiler (FCProfiler), a deterministic

pipeline for recipe-level nutrient estimation. The approach decomposes the task into explicit stages: unit canonicalization following standardized measurement rules such as NIST SP 811 (Thompson and Taylor, 2008), ingredient-specific gram conversion grounded in USDA FoodData Central (U.S. Department of Agriculture, Agricultural Research Service, Beltsville Human Nutrition Research Center, 2026), and structured nutrient estimation. To improve coverage of informal and diverse measurement expressions, unit canonicalization is further expanded using corpus-driven mining from Recipe1M+ (Marin et al., 2021).

This structured design leads to the following key contributions:

- We design a deterministic and interpretable pipeline for recipe-level nutrient estimation, explicitly separating unit canonicalization, gram conversion, and nutrient estimation.
- We demonstrate through ablation studies that reliable unit canonicalization and ingredient-specific gram conversion are the primary factors driving nutrient estimation accuracy on FoodBench-QA dataset (Eftimov et al., 2025b).
- We show that the proposed pipeline attains performance competitive with FoodyLLM on FoodBench dataset through explicit measurement grounding (Eftimov et al., 2025a).

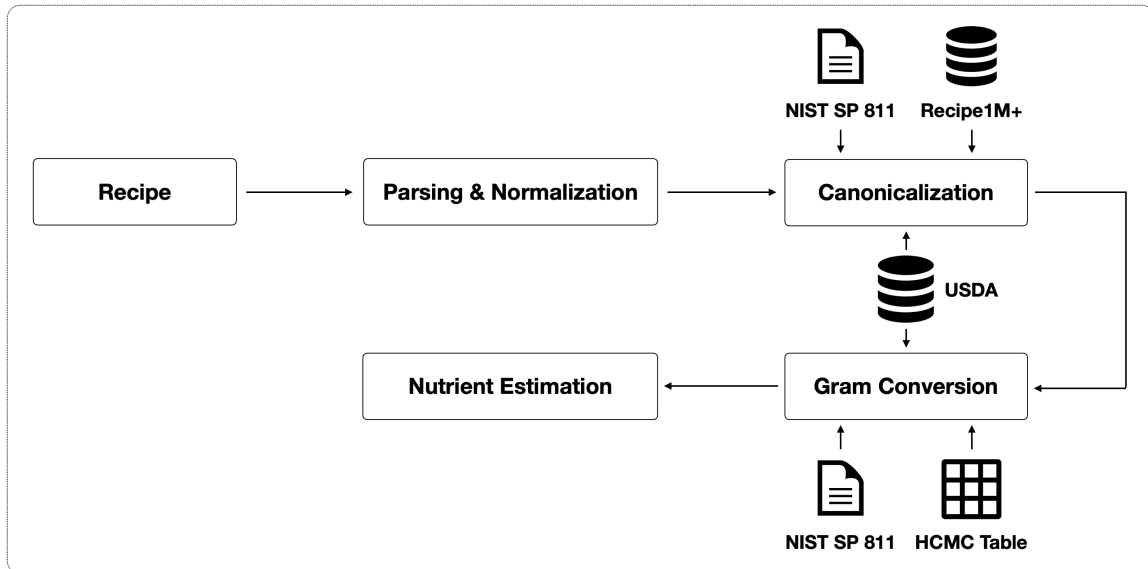


Figure 1: FCProfiler pipeline overview

## 2. Related Work

Prior work on food-related nutrient estimation has been explored through various paradigms, including image-based and text-based approaches, as well as recent LLM-based and multimodal approaches.

Image-based approaches to nutrient estimation involve food image segmentation, portion size estimation, volume-to-mass conversion, and calorie estimation (Pouladzadeh et al., 2014; Myers et al., 2015). Nutrient values can be derived through database lookup based on food type and mass. However, these approaches primarily focus on calorie estimation and rely on intermediate physical approximations, making them sensitive to error propagation and limiting their applicability to accurate nutrient estimation.

Text-based approaches have explored predicting nutrient values directly from textual descriptions. P-NUT demonstrated that macronutrient content can be predicted from short food descriptions using text embeddings and clustering-based regression models (Ispirova et al., 2020). However, this approach relies on limited textual descriptions and does not incorporate ingredient-level structure, quantity information, or explicit grounding in food composition databases, reducing its applicability to realistic recipe scenarios. To address these limitations, subsequent studies extended the problem to structured recipe data (Ispirova et al., 2022), extracting ingredients, quantities, and measurement units using rule-based and dictionary-based methods and mapping them to food composition databases such as USDA FoodData Central. While this enables ingredient-aware representations and incorporation of measurement information, nutrient estimation is

still treated as a learned prediction task rather than an explicit measurement resolution process.

Recent approaches have adopted LLMs and multimodal approaches to address multiple food-related tasks in a unified framework. FoodyLLM integrates nutrient estimation, traffic-light classification, and food entity linking as a multi-task instruction learning problem (Gjorgjevikj et al., 2026), while recent multimodal models further incorporate visual and textual inputs for holistic food understanding (Yin et al., 2025). However, these approaches rely on implicit reasoning within end-to-end architectures and do not provide explicit control over intermediate steps such as unit normalization and ingredient-specific gram conversion, which are critical for reproducible and numerically consistent nutrient estimation.

In contrast to these approaches, FCProfiler formulates recipe-level nutrient estimation as a structured measurement resolution framework. Rather than predicting nutrient values directly, the proposed pipeline explicitly models ingredient parsing, unit canonicalization, gram conversion, and database-grounded computation, enabling interpretable, reproducible, and numerically traceable nutrient estimation.

## 3. Methodology

FCProfiler is decomposed into structured stages: ingredient parsing and normalization, unit canonicalization, gram conversion, and nutrient estimation. Each stage produces explicit intermediate representations, enabling traceability from raw ingredient text to final nutrient values.

| Dataset                    | Split | Protein | Sugars | Fat    | Saturates |
|----------------------------|-------|---------|--------|--------|-----------|
| <i>Ingredients Only</i>    | Train | 0.9511  | 0.9065 | 0.9146 | 0.9291    |
| <i>Ingredients Only</i>    | Test  | 0.9461  | 0.9200 | 0.9224 | 0.9336    |
| <i>Title + Ingredients</i> | Train | 0.9507  | 0.9070 | 0.9142 | 0.9294    |
| <i>Title + Ingredients</i> | Test  | 0.9461  | 0.9178 | 0.9236 | 0.9325    |

Table 1: Nutrient estimation accuracy of FCProfiler on FoodBench-QA train and test datasets under tolerance-based evaluation protocol

| Dataset                    | Model      | Protein              | Sugars               | Fat                  | Saturates            |
|----------------------------|------------|----------------------|----------------------|----------------------|----------------------|
| <i>Ingredients Only</i>    | FoodyLLM   | <b>0.972 ± 0.002</b> | 0.917 ± 0.005        | 0.920 ± 0.004        | <b>0.938 ± 0.004</b> |
| <i>Ingredients Only</i>    | FCProfiler | 0.945 ± 0.002        | <b>0.920 ± 0.003</b> | <b>0.922 ± 0.002</b> | 0.934 ± 0.001        |
| <i>Title + Ingredients</i> | FoodyLLM   | <b>0.967 ± 0.002</b> | 0.915 ± 0.008        | 0.912 ± 0.003        | <b>0.934 ± 0.004</b> |
| <i>Title + Ingredients</i> | FCProfiler | 0.946 ± 0.002        | <b>0.919 ± 0.003</b> | <b>0.924 ± 0.002</b> | 0.933 ± 0.002        |

Table 2: Comparative evaluation of FCProfiler and FoodyLLM on FoodBench test dataset under tolerance-based evaluation protocol

### 3.1. Ingredient Parsing & Normalization

Given an input recipe, FCProfiler extracts structured tuples consisting of an ingredient name, quantity, and measurement unit. Ingredient parsing is performed using a rule-based approach based on regular expressions and token-level segmentation. The extracted ingredient names are then normalized through string-level preprocessing to produce representations for subsequent stages.

### 3.2. Unit Canonicalization

Household measurement expressions exhibit substantial variability (e.g., “tbsp”, “T”, “fl. oz”, “pinch”). To standardize these forms, FCProfiler applies layered canonicalization combining standardized measurement conventions (e.g., NIST SP 811), unit aliases derived from USDA FoodData Central, and corpus-driven unit expansion from Recipe1M+. This process maps heterogeneous expressions to consistent unit representations compatible with nutrient databases.

### 3.3. Gram Conversion

Ingredient quantities are converted to grams using a prioritized conversion strategy. Direct SI units are deterministically mapped. For volume- and count-based units, conversion is performed using ingredient-specific portion information from USDA FoodData Central or standardized mappings from Household Cooking Measurements Conversion (HCMC) table, ensuring ingredient-specific mass estimation rather than relying on global density assumptions.

### 3.4. Nutrient Estimation

After gram conversion, nutrient values are retrieved from USDA FoodData Central and aggregated at

the recipe level. Final values are normalized per 100g following the shared task protocol. All computations are deterministic and database-grounded.

## 4. Experimental Results

### 4.1. Datasets

Experiments are conducted on FoodBench-QA dataset, which consists of English recipe text, evaluated under EU Regulation 1169/2011 tolerance standards. FoodBench-QA dataset provides two derived datasets for nutrient estimation task: *Ingredients Only* and *Title + Ingredients*. Each dataset includes predefined train and test splits, and evaluation is performed using tolerance-based nutrient accuracy for protein, sugars, fat, and saturates.

For comparison with FoodyLLM, we additionally evaluate on FoodBench dataset, which follows the same subset construction with *Ingredients Only* and *Title + Ingredients* and the same tolerance-based evaluation protocol for all nutrients. The proposed pipeline is evaluated using only the provided test splits from five independently sampled datasets.

Standardized measurement guidelines (NIST SP 811) and food-related resources, including USDA FoodData Central, Recipe1M+, and HCMC, are employed throughout unit canonicalization and gram conversion.

### 4.2. Results

Table 1 represents the final tolerance-based nutrient accuracy of FCProfiler under both datasets: *Ingredients Only* and *Title + Ingredients*. Across both train and test splits, FCProfiler achieves consistently high accuracy for all nutrients. On the test split, the full configuration with canonicalization and layered gram conversion reaches 0.9461 for pro-

| Unit Canonicalization Strategy | Split | Protein       | Sugars        | Fat           | Saturates     |
|--------------------------------|-------|---------------|---------------|---------------|---------------|
| None                           | Train | 0.5378        | 0.4259        | 0.3251        | 0.3935        |
| NIST SP                        | Train | 0.9332        | 0.8843        | 0.8879        | 0.9088        |
| NIST SP + USDA                 | Train | 0.9441        | 0.8981        | 0.9032        | 0.9201        |
| NIST SP + USDA + Recipe1M+     | Train | <b>0.9511</b> | <b>0.9065</b> | <b>0.9146</b> | <b>0.9291</b> |
| None                           | Test  | 0.5039        | 0.4045        | 0.2893        | 0.3538        |
| NIST SP                        | Test  | 0.9404        | 0.9134        | 0.9146        | 0.9287        |
| NIST SP + USDA                 | Test  | 0.9431        | 0.9152        | 0.9170        | 0.9295        |
| NIST SP + USDA + Recipe1M+     | Test  | <b>0.9461</b> | <b>0.9200</b> | <b>0.9224</b> | <b>0.9336</b> |

Table 3: Ablation results of unit canonicalization strategies for nutrient estimation

| Gram Conversion Strategy        | Split | Protein       | Sugars        | Fat           | Saturates     |
|---------------------------------|-------|---------------|---------------|---------------|---------------|
| Direct conversion only          | Train | 0.6282        | 0.5128        | 0.4252        | 0.5088        |
| + USDA portion-based conversion | Train | 0.9320        | 0.8901        | 0.8959        | 0.9140        |
| + HCMC table-based conversion   | Train | 0.9441        | 0.8981        | 0.9032        | 0.9201        |
| + Heuristic fallback            | Train | <b>0.9511</b> | <b>0.9065</b> | <b>0.9146</b> | <b>0.9291</b> |
| Direct conversion only          | Test  | 0.5795        | 0.4717        | 0.3672        | 0.4550        |
| + USDA portion-based conversion | Test  | 0.9285        | 0.9036        | 0.9073        | 0.9227        |
| + HCMC table-based conversion   | Test  | 0.9431        | 0.9152        | 0.9170        | 0.9295        |
| + Heuristic fallback            | Test  | <b>0.9461</b> | <b>0.9200</b> | <b>0.9224</b> | <b>0.9336</b> |

Table 4: Ablation results of gram conversion strategies for nutrient estimation

tein, 0.9200 for sugars, 0.9224 for fat, and 0.9336 for saturates.

### 4.3. Comparison with FoodyLLM

Table 2 compares FCProfiler with FoodyLLM on FoodBench test dataset. FCProfiler achieves comparable performance across all nutrients for both *Ingredients Only* and *Title + Ingredients* datasets. In particular, FCProfiler shows slightly higher accuracy on sugars and fat across both datasets. The reported results are expressed as mean and standard deviation over five independently sampled datasets, and FoodyLLM results are taken directly from the original paper. These results indicate that explicit measurement grounding can provide a competitive alternative to end-to-end neural reasoning for nutrient estimation.

### 4.4. Ablation Study

Tables 3 and 4 represent ablation results on *Ingredients Only* dataset of FoodBench-QA dataset.

#### 4.4.1. Unit Canonicalization

When unit canonicalization is completely disabled, performance collapses across all nutrients, particularly on the test split. This confirms that resolving heterogeneous household measurement expressions is essential for meaningful nutrient estimation.

Applying NIST SP-based canonicalization alone produces a substantial improvement, with test accuracy above 0.91 for all nutrients. Incorporating USDA alias expansion and corpus-driven

Recipe1M+ unit expansion yields further improvements, with the full configuration achieving the best results across all nutrients.

The most significant improvements are observed when transitioning from no canonicalization to NIST SP-based canonicalization, indicating that standardizing measurement representations is a key factor for tolerance-aware accuracy.

#### 4.4.2. Gram Conversion

Using direct unit-to-gram conversion alone substantially reduces performance on the test split, with protein accuracy dropping to 0.5795 and fat to 0.3672. This indicates that direct conversion alone is insufficient for volume- and count-based units commonly found in recipes.

Incorporating USDA portion-based conversion significantly improves performance, raising test accuracy to 0.9285 for protein and 0.9073 for fat. Adding HCMC table-based conversion further increases protein accuracy to 0.9431 and fat to 0.9170. Finally, heuristic fallback mechanisms yield the best results, achieving 0.9461 for protein and 0.9224 for fat on the test split.

The combined strategy achieves the highest accuracy across all nutrients, indicating that nutrient estimation accuracy largely depends on accurate mass grounding.

## 5. Discussion

Accurate nutrient estimation requires resolving ingredient quantities into standardized gram values



through ingredient-specific measurement interpretation. For instance, in “1 teaspoon wheat flour,” the unit must be interpreted and converted into grams based on ingredient-specific portion information. In FCProfiler, this is resolved by mapping the unit to *tsp* and grounding it in USDA portion data, yielding 2.60g. This example underscores the importance of accurate unit canonicalization and ingredient-specific gram conversion for reliable nutrient estimation.

This observation extends consistently across ingredients. The same unit, such as teaspoon or cup, corresponds to different masses depending on the ingredient type, making fixed density assumptions inherently unreliable. By grounding conversions in ingredient-specific portion metadata, FCProfiler reduces systematic errors introduced by generic conversion rules.

Unit canonicalization is also essential as many recipe measurements are expressed in heterogeneous and non-standardized forms. The results in Table 3 show that NIST SP-based canonicalization alone leads to substantial performance improvements, indicating that mapping diverse measurement expressions into a consistent unit system is important for reliable gram conversion. Additional alias expansion and corpus-driven expansion provide incremental improvements, with their contribution expected to be more pronounced in more heterogeneous recipe corpora.

## 6. Conclusion

We propose FCProfiler, a deterministic and interpretable pipeline for recipe-level nutrient estimation. Rather than relying on generative models, the pipeline performs structured measurement processing and database-grounded computation. Experimental results indicate that unit canonicalization and ingredient-specific gram conversion are critical to nutrient estimation accuracy. In particular, portion-based grounding using USDA FoodData Central enables realistic volume-to-mass conversion without relying on global density assumptions, while preserving traceability throughout the pipeline. FCProfiler also achieves comparable performance to FoodyLLM on FoodBench dataset, indicating the effectiveness of explicit measurement grounding.

These findings indicate that reliable food and nutrition question answering depends on accurate unit canonicalization, ingredient-specific gram conversion, and database grounding. The source code is publicly available at <https://github.com/rbls-lab/FCProfiler> to support reproducibility, and includes implementation details and experimental configurations.

## 7. Limitations

Despite strong performance, portion ambiguity remains a practical limitation. Expressions such as “1 pack noodles” or “1 serving sauce” often lack standardized gram mappings. Although heuristic fallback strategies provide approximate mass values, these approximations introduce uncertainty in nutrient computation.

In the future, we would like to apply food category-aware constraints derived from nutrient distributions within USDA FoodData Central. By modeling expected nutrient ranges per food category, it may be possible to detect and adjust implausible computations caused by portion-related mass conversion errors while maintaining interpretability.

The same deterministic pipeline of FCProfiler can be extended to other FoodBench-QA shared tasks, including FSA traffic-light prediction and ontology-grounded food entity linking. This indicates that structured measurement resolution and database grounding provide a generalizable basis for food and nutrition question answering.

## 8. Acknowledgements

We are grateful to Dongwoo Kim from Korea National Open University for his insightful feedback, and to Kyunghye An of Seongseo Public Health Subcenter and Sunok Kim of Soonok Bakery for their helpful comments and practical insights. We also thank the anonymous reviewers for their valuable comments.

## 9. Ethics Statement

FCProfiler is evaluated on publicly available recipe data and uses open nutritional resources. The study does not involve personal or sensitive information. Nutrient estimates are database-grounded approximations and should not be interpreted as medical advice.

## 10. Bibliographical References

- Lorenzo Bairati. 2017. [The food consumer’s right to information on product country of origin: trends and outlook, beyond eu regulation 1169/2011](#). *Journal of European Consumer and Market Law*, 6(1):9–16.
- Ana Gjorgjevikj, Matej Martinc, Gjorgjina Cenikj, Riste Stojanov, Jan Drole, Gordana Ispirova, Giulia Menichetti, Nives Ogrinc, Dimitar Trajanov, Sašo Džeroski, Barbara Koroušić Seljak, and Tome Eftimov. 2026. [Large language models in](#)

- food and nutrition science: Opportunities, challenges, and the case of foodylm. *Current Research in Food Science*, 12:101351.
- Gordana Ispirova, Tome Eftimov, and Barbara Koroušić Seljak. 2020. [P-nut: Predicting nutrient content from short text descriptions](#). *Mathematics*, 8(10).
- Gordana Ispirova, Tome Eftimov, and Barbara Koroušić Seljak. 2022. [Predefined domain specific embeddings of food concepts and recipes: A case study on heterogeneous recipe datasets](#). In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4074–4083.
- Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. [A survey on food computing](#). *ACM Computing Surveys*, 52(5).
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. [Large language models: A survey](#). ArXiv preprint arXiv:2402.06196.
- Austin Myers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin Murphy. 2015. [Im2calories: towards an automated mobile vision food diary](#). In *ICCV*.
- Parisa Pouladzadeh, Shervin Shirmohammadi, and Rana Al-Maghrabi. 2014. [Measuring calorie and nutrition from food image](#). *IEEE Transactions on Instrumentation and Measurement*, 63(8):1947–1956.
- Ambler Thompson and Barry Taylor. 2008. [Guide for the use of the international system of units \(si\)](#). Special Publication (NIST SP), National Institute of Standards and Technology.
- Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. 2025. [Foodlmm: A versatile food assistant using large multi-modal model](#). *IEEE Transactions on Multimedia*, 27:6949–6961.
- Eftimov, Tome and Gjorgjevikj, Ana and Martinc, Matej and Cenikj, Gjorgjina and Džeroski, Sašo and Koroušić Seljak, Barbara. 2025b. [FoodBench-QA: Shared Task on Grounded Food & Nutrition Question Answering](#). Accessed: 2026-02-18.
- Marin, Javier and Biswas, Aritro and Ofli, Ferda and Hynes, Nicholas and Salvador, Amaia and Aytar, Yusuf and Weber, Ingmar and Torralba, Antonio. 2021. [Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images](#).
- U.S. Department of Agriculture, Agricultural Research Service, Beltsville Human Nutrition Research Center. 2026. [FoodData Central](#). Accessed: 2026-02-18.

## 11. Language Resource References

- Eftimov, Tome and Gjorgjevikj, Ana and Martinc, Matej and Cenikj, Gjorgjina and Darko, Sasan-ski and Stojanov, Riste and Džeroski, Sašo and Koroušić Seljak, Barbara. 2025a. [FoodBench: A Domain-Targeted Question-Answering Dataset for Food and Nutrition Applications](#). Zenodo.

# DocUA at CRF Filling 2026: LLM StructCore — Schema-Guided Reasoning Condensation and Deterministic Compilation

Serhii Zabolotnii

Cherkasy State Business College

Cherkasy, Ukraine

zabolotnii.serhii@csbc.edu.ua

## Abstract

Automatically filling Case Report Forms (CRFs) from clinical notes is challenging due to noisy language, strict output contracts, and the high cost of false positives. We describe our CL4Health 2026 submission for Dyspnea CRF filling (134 items) using a *contract-driven* two-stage design grounded in *Schema-Guided Reasoning* (SGR) (Abdullin, 2026). The key task property is **extreme sparsity**: the majority of fields are `unknown`, and official scoring penalizes both empty values and unsupported predictions. We shift from a single-step “LLM predicts 134 fields” approach to a decomposition where (i) Stage 1 produces a stable SGR-style JSON summary with exactly 9 domain keys, and (ii) Stage 2 is a fully deterministic, 0-LLM compiler that parses the Stage 1 summary, canonicalizes item names (optionally using a UMLS alias map with 134/134 coverage), normalizes predictions to the official controlled vocabulary (13 categories), applies evidence-gated false-positive filters, and expands the output into the required 134-item format. On the dev80 split, the best teacher configuration (Mistral Large 3 Stage 1 → Stage 2 deterministic) achieves macro-F1 **0.6543** (EN) and **0.6905** (IT); on the hidden test200, the submitted English variant scores **0.63** on Codabench. The pipeline is language-agnostic: Italian results match or exceed English with no language-specific engineering.

**Keywords:** clinical NLP, information extraction, schema-guided reasoning, controlled vocabulary, reproducibility, edge AI

## 1. Introduction

The CL4Health 2026 CRF filling shared task requires mapping unstructured clinical notes to a strict 134-item Dyspnea CRF. In practice, end-to-end LLM extraction often fails either due to output format drift (invalid JSON / JSONL structure) or due to unsupported “hallucinated” fills that increase false positives (FP). Our design goal is to separate **(a)** clinical information condensation from **(b)** contract enforcement and controlled-vocabulary normalization, while using SGR to make the intermediate representation stable and machine-checkable.

**Schema-Guided Reasoning (SGR).** SGR (Abdullin, 2026) is an architectural pattern that constrains an LLM to produce typed, schema-conformant outputs via structured output or constrained decoding (Dong et al., 2024; Willard and Louf, 2023), turning domain knowledge into executable data contracts. This reduces variability and improves reliability compared to free-form text prompting. In our system, Stage 1 uses an SGR-style schema with three core patterns to produce a stable 9-key domain summary:

- **Cascade (Domain Scaffold):** The LLM follows a fixed sequence of 9 clinical domains (Demographics → Vitals → Labs → Problems → Symptoms → Medications → Procedures → Utilization → Disposition) to ensure uniform

coverage of the clinical narrative.

- **Cycle (Checklists):** The prompt enforces implicit checklists within each domain (e.g., examining 28 specific CRF problems, all critical vitals) to improve recall on sparse items.
- **Routing (Vocabulary):** While Stage 1 emits unconstrained text strings within the JSON structure, Stage 2 deterministically routes each prediction into one of 13 required controlled-vocabulary categories (Table 1).

Stage 2 then deterministically compiles this summary into the official CRF submission format, ensuring 100% structural compliance.

## 2. Task and Data

### 2.1. Datasets and Splits

We use the official Hugging Face datasets (Ferrazzi et al., 2025) (NLP-FBK/dyspnea-crf-\*) and, for teacher generation, the in-domain unannotated clinical notes set (NLP-FBK/dyspnea-clinical-notes, 2,667 records from SGB hospital, Italy). Train contains 10 annotated records, development 80, and test 200 (hidden ground truth).

A key discovery is **extreme sparsity**: across the development set, the median number of non-`unknown` items per document is only  $\sim 12$  out of 134 total. The most frequent non-unknown items are in VITALS (*heart rate*, *blood pressure*, *spo2*)

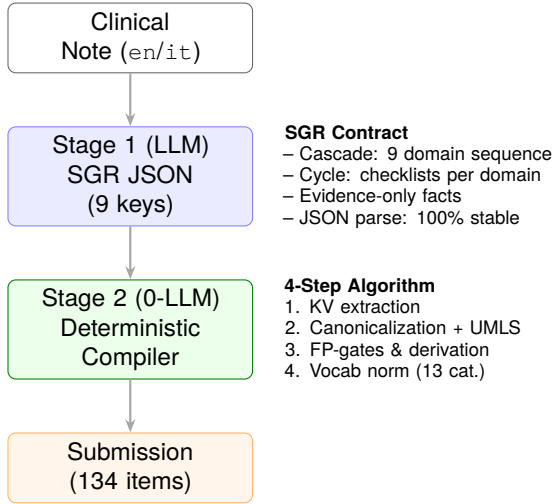


Figure 1: The LLM StructCore two-stage pipeline architecture. Stage 1 condenses the clinical note into a 9-key SGR JSON; Stage 2 deterministically compiles it into the 134-item submission.

and LABS (*hemoglobin, creatinine*), while many diagnosis and procedure items appear in <5% of documents. This sparsity means that `unknown` is by far the most common class, and correctly predicting `unknown` for absent items is as important as correctly extracting present values. The official scorer (Ferrazzi et al., 2026b) uses macro-F1 over items, which heavily penalizes both false positives (FP: predicting a value for an `unknown` item) and false negatives (FN: predicting `unknown` for a present item).

## 2.2. Contract-Driven Dataset Parsing

We align our pipeline contracts directly to the dataset schema: (i) the ordered list of 134 items is inferred from `annotations[*].item` and reused end-to-end; (ii) document identifiers are normalized to a plain ID (prefix before the first underscore) for internal alignment, and suffixed with the language tag (`_en/_it`) only for final submissions; (iii) intermediate representations are sparse, but the final builder always expands to all 134 items, filling missing values as `unknown`.

## 3. System Overview

Our pipeline follows a strict two-stage decomposition:

```

Clinical Note → Stage1 (LLM, SGR JSON) → Stage2 (O-LLM compiler) → Submission.jsonl

```

### 3.1. Stage 1: SGR-Style JSON Summary

Stage 1 produces a single JSON object with exactly 9 domain keys: `DEMOGRAPHICS`, `VITALS`, `LABS`, `PROBLEMS`, `SYMPTOMS`, `MEDICATIONS`, `PROCEDURES`, `UTILIZATION`, `DISPOSITION`.

The summary is designed to be **sparse** (include only evidence-supported facts) and **format-stable** (JSON parse success on every input). On the train10 split, we observe `json_parse_ok=10/10` and `thinking_leak=0/10` even on quantized 4B models served via `llama.cpp` without grammar-constrained decoding.

**Prompt Structure.** The Stage 1 prompt instructs the model to produce a single JSON object following a strict template. Each of the 9 keys maps to a free-text string summarizing the relevant clinical domain. The prompt includes: (i) explicit definitions of each key’s scope (e.g., `PROBLEMS` covers past medical history, `SYMPTOMS` covers presenting complaints); (ii) a checklist of high-priority items per domain (e.g., 28 specific problems listed in the CRF ontology, all vital signs, key laboratory values); (iii) instructions to use `Key: Value` formatting within each string for machine-parseable extraction; (iv) a sparsity directive: “include only facts explicitly stated in the note; do not infer or guess”. This structured prompt design implements the SGR Cascade and Cycle patterns, ensuring systematic coverage while preventing hallucinated fills.

**Format Stability Without Constrained Decoding.** A notable finding is that our SGR prompt achieves 100% JSON parse success *without* requiring grammar-constrained decoding (e.g., XGrammar (Dong et al., 2024) or Outlines (Willard and Louf, 2023)). We attribute this to the simplicity of the target schema (a flat 9-key object with string values) and the explicit formatting instructions. This makes the approach compatible with any inference backend (`llama.cpp`, `vLLM`, OpenAI-compatible APIs, Gemini Vertex) without backend-specific grammar support.

**Multi-Slice Input Strategy.** Long clinical notes may exceed the model’s effective context window. We implement a **multi-slice** strategy: the note is split into overlapping windows (e.g., `middle+head_tail`, window size  $w=4$ ), each window is summarized independently via Stage 1, and the resulting JSON objects are merged with a last-writer-wins policy per key. For the `PROBLEMS`, `MEDICATIONS`, and `LABS` keys (which accumulate multiple facts), we concatenate the strings from all windows rather than overwriting. This preserves recall



for facts spread across the note while keeping each individual inference within context bounds.

**Teacher Data Generation.** For the Stage 1 teacher pipeline, we use large API-accessible models to produce high-quality SGR summaries. We tested three teacher configurations:

- **Mistral Large 3** (675B, via NVIDIA API) — best overall quality, supporting `response_format=json_object` on most NVIDIA endpoints (with auto-fallback to parser-based extraction for Mistral-family models where structured output returns 400 errors).
- **Qwen 3.5** (397B, via NVIDIA API) — comparable quality but higher FP; requires `enable_thinking=false` to prevent “Thinking Process” leakage into the output.
- **Claude Sonnet 4.6** (via Anthropic API) — recall-optimized prompts achieve strong results but not submitted due to participation limits.

Teacher-generated summaries are compiled through the same deterministic Stage 2, enabling precise ablation of Stage 1 quality vs. Stage 2 deterministic logic.

**Concrete Example.** Given a Stage 1 summary JSON with VITALS: "Heart Rate: 105 bpm (tachycardic). Blood Pressure: 90/60 mmHg (hypotensive). SpO2: 88%. Respiratory Rate: 28/min (tachypneic).", the Stage 2 compiler:

1. Extracts KV pairs: `{heart rate: "105 bpm (tachycardic)", blood pressure: "90/60 mmHg (hypotensive)", spo2: "88%", respiratory rate: "28/min (tachypneic)"}`.
2. Canonicalizes: all keys match the official item list.
3. No FP-gates triggered for vital signs.
4. Normalizes: `heart rate` → `tachycardic`, `blood pressure` → `hypotensive`, `spo2` → `88`, `respiratory rate` → `tachypneic`.

The remaining 130 items are filled as `unknown`.

### 3.2. Stage 2: Deterministic Compiler (O-LLM)

Stage 2 performs **no LLM calls**. It implements a 4-step deterministic algorithm (Algorithm 3.2):

**Step 1: KV Extraction.** Parse the Stage 1 summary text into a flat dictionary of `(item_name, value_string)` pairs by matching Key: Value lines in each domain section. Lines with `not stated` or `unknown` values are skipped.

**Step 2: Canonicalization.** Normalize item names via: (a) exact match against the official

| #                      | Vocabulary Category                 | Items      |
|------------------------|-------------------------------------|------------|
| 1                      | chronic (4-class)                   | 7          |
| 2                      | neoplasia (3-class)                 | 1          |
| 3                      | duration (short/long)               | 2          |
| 4                      | binary (y/n)                        | 61         |
| 5                      | mobility (4-class)                  | 1          |
| 6                      | consciousness (AVPU)                | 1          |
| 7                      | respiratory rate (3-cat)            | 1          |
| 8                      | body temperature (3-cat)            | 1          |
| 9                      | heart rate (3-cat)                  | 1          |
| 10                     | blood pressure (3-cat)              | 1          |
| 11                     | respiratory distress (current/past) | 1          |
| 12                     | diagnostic tests (pos/neg)          | 4          |
| 13                     | labs (measured/numeric)             | 25         |
| <b>Total</b>           |                                     | <b>107</b> |
| Unknown-only remainder |                                     | 27         |
| <b>Grand Total</b>     |                                     | <b>134</b> |

Table 1: The 13 controlled-vocabulary categories defined by the CRF ontology. Stage 2 maps every extracted value string into exactly one of these categories.

134-item list; (b) case-folding and punctuation stripping (e.g., `first episode` → `first episod`); (c) abbreviation expansion (70+ rules, e.g., `hr` → `heart rate`, `cxr` → `chest rx`); (d) optionally, UMLS alias resolution using a curated map with **134/134 item coverage** (Aronson, 2001).

**Step 3: FP-Gates & Derivation.** Derive high-FN composite items from Stage 1 evidence: *poly-pharmacological therapy* is inferred when  $\geq 8$  distinct medications are listed; *antihypertensive therapy* is detected by matching 35 antihypertensive agents; *cardiovascular diseases* is derived from therapy flags or PMH keywords. Then, strict regex-based evidence gates **drop** unsupported positive predictions for items with high FP cost: *active neoplasia* (requires cancer vocabulary + activity cues), *arrhythmia* (requires specific rhythm terms, blocks on `NSR`), *acute coronary syndrome* (requires ACS vocabulary + supporting evidence), and *presence of dyspnea* (blocks on explicit negation).

**Step 4: Vocabulary Normalization (13 Categories).** Map each value string to the official controlled vocabulary (Table 1): `binary` → `y/n`; `AVPU` → `A/V/P`; `chronic` items → `4-class` scale; `neoplasia` → `3-class` activity; `duration` → `short/long`; `vitals` → `categorical` (e.g., `tachycardic/normocardic/bradycardic`); `labs` → `numeric` extraction or `measured`. Finally, expand to all 134 items, filling absent items as `unknown`.



### 3.3. Optional UMLS Alias Mapping

We curated a UMLS-based alias map for all 134 CRF items (**coverage 134/134**) and integrated it as an optional pipeline mode (`--use-umls-mapping`) to reduce key-name brittleness in model outputs. The map associates each official item name with its UMLS CUI and preferred name, plus manually verified synonyms. This mode is activated in all final submissions.

## 4. Experiments and Results

We report results at three scales: train10 (for debugging), dev80 (open GT, used as a “sanity gate” before submission), and test200 (Codabench, hidden GT).

### 4.1. Development Results (dev80, Open GT)

Table 2 summarizes our main dev80 runs across both English and Italian. All use the same deterministic Stage 2 compiler with UMLS mapping and FP-gates; EN runs use `rules_v3a`, while the latest IT variant uses the improved `rules_v3e`.

**Key observations.** (1) Mistral Large 3 and Qwen 3.5 produce comparable TP counts (236) on EN, but Qwen generates more FP (211 vs. 170), likely due to less conservative summarization. (2) The FN count is identical (128) across both EN teacher models, indicating that Stage 2’s deterministic derivation rules have a ceiling that is *Stage 1 recall-limited*. (3) On train10, teacher Stage 1 (Gemini) dramatically reduces FN (9 vs. 22 for student) but increases FP (35 vs. 13), motivating our evidence-gated FP filters. (4) The student model (MedGemma 4B, local) achieves  $F1=0.6521$  on train10 with very low FP (13), suggesting that smaller models produce more conservative summaries. (5) Italian dev80 achieves higher  $F1$  (0.6905) than English (0.6543) with more TP (266 vs. 236) and fewer FN (106 vs. 128), suggesting that the original Italian clinical notes are more structured and easier for Stage 1 extraction. The `v3e` rule set reduces 3 FP compared to `v3a` (198 vs. 201) through tighter FP-drop gates on *chest pain*, *heart failure*, and *arrhythmia*.

**Multi-slice input ablation.** Both dev80 teacher runs use the `middle+head_tail` multi-slice strategy with  $w=4$ . Compared to single-window runs on the same models, multi-slice recovers +12–18 additional TP at the cost of +8–15 FP, a net positive trade. The improvement is concentrated in the LABS and MEDICATIONS domains, where relevant information is often scattered across the note.

### 4.2. Test Results (Codabench)

Table 3 shows the submitted and evaluated test200 results, alongside the leaderboard context.

The best submitted English system achieves  $F1 = 0.63$  using Mistral Large 3 as the Stage 1 teacher, placing within 0.05 of the top-ranking result ( $F1 = 0.68$ ). The Anthropic variant (Claude Sonnet 4.6 with recall-optimized prompts) produces comparable quality ( $F1 = 0.62$ ) but was not submitted as the final entry due to the participation limit. The Italian variant using `rules_v3e` was fully prepared (dev80 IT  $F1 = 0.6905$ ) but could not be submitted as all attempt slots were exhausted.

Our findings suggest that large open-weight models like Mistral produce Stage 1 summaries competitive with larger proprietary systems.

**Dev10 Cross-Language Validation.** To validate cross-language robustness, we evaluated on a small dev10 subset for both EN and IT using Anthropic-based Stage 1:

The Italian results are notably higher than English on this small dev10 sample. This trend is **confirmed at scale on dev80**: IT achieves  $F1 = 0.6905$  vs. EN = 0.6543 (Table 2). We attribute this to two factors: (1) the Italian clinical notes from SGB hospital are written in a more structured, formulaic style than their English translations, facilitating cleaner Stage 1 extraction; (2) certain CRF items use Italian medical terminology that maps more directly to the ontology (e.g., *dispnea* → *presence of dyspnea*).

### 4.3. Error Analysis

**FP-gates effectiveness.** The FP-gate mechanism (Step 3 of Stage 2) drops an average of 3.2 items per document that would otherwise be false positives. Table 5 shows the per-item impact of FP-gates on dev80.

The most impactful gate is *active neoplasia*: without evidence gating, benign mentions of “cancer screening” or “family history of cancer” produce FP; the gate requires explicit cancer vocabulary *plus* activity cues (e.g., “chemotherapy”, “progression”, “untreated”). For *arrhythmia*, the gate blocks positive predictions when the evidence contains Normal Sinus Rhythm (NSR) mentions, preventing common misclassification of ECG-described patients.

**Italian-specific error patterns (dev80 IT).** Detailed item-level diagnostics on the dev80 IT split (Mistral Large 3 → `rules_v3e`) reveal both shared and language-specific error patterns. The top FN sources are: *foreign body in the airways* (11 FN, Stage 1 never surfaces this item), *presence of dyspnea* (7 FN / 8 FP — tension between coverage and FP), *diffuse vascular disease* (7 FN / 6 FP), and *agitation* (7 FN, not extracted by Stage 1). The top

| Lang                                      | Stage 1 Model              | F1            | TP  | FP  | FN  |
|-------------------------------------------|----------------------------|---------------|-----|-----|-----|
| EN                                        | Mistral Large 3 (675B)     | 0.6543        | 236 | 170 | 128 |
| IT                                        | Mistral Large 3 (v3e)      | <b>0.6905</b> | 266 | 198 | 106 |
| IT                                        | Mistral Large 3 (v3a)      | 0.6900        | 266 | 201 | 106 |
| EN                                        | Qwen 3.5 (397B, think=off) | 0.6411        | 236 | 211 | 128 |
| <i>Smaller-scale ablations (train10):</i> |                            |               |     |     |     |
| EN                                        | Gemini Flash (teacher)     | 0.6763        | 68  | 35  | 9   |
| EN                                        | MedGemma 4B (student)      | 0.6521        | 57  | 13  | 22  |
| EN                                        | SGR-only baseline          | 0.4384        | 39  | 44  | 39  |

Table 2: Main results on dev80 (N=80; top rows, both languages) and ablations on train10 (N=10; bottom rows). All variants use the same deterministic Stage 2 compiler. *v3e* denotes the latest Stage 2 rule set with improved FP-drop gates for Italian. F1 = official macro-F1 scorer; TP/FP/FN = item-level counts.

| System Variant              | F1   | Lang | Team  | Status                 |
|-----------------------------|------|------|-------|------------------------|
| Mistral L3 → Stage2(det)    | 0.63 | EN   | DocUA | Submitted              |
| No FP-gates variant         | 0.61 | EN   | DocUA | Submitted              |
| Claude Sonnet → Stage2(det) | 0.62 | EN   | DocUA | Evaluated <sup>†</sup> |
| Mistral L3 recall (v3e)     | —    | IT   | DocUA | Prepared <sup>‡</sup>  |

Table 3: Test200 results on Codabench. All DocUA variants use deterministic Stage 2 + UMLS mapping. <sup>†</sup> Evaluated but not submitted as final entry due to participation limits. <sup>‡</sup> Italian variant prepared (Stage 1 SGR9 recall, chars=8000, multi-slice, *rules\_v3e*) but not submitted due to exhausted attempt slots.

| Language     | F1            | TP | FP | FN |
|--------------|---------------|----|----|----|
| Italian (IT) | <b>0.9000</b> | 59 | 9  | 4  |
| English (EN) | 0.8115        | 55 | 18 | 7  |

Table 4: Dev10 cross-language validation (Anthropic recall Stage 1 → Stage 2 det).

| Gated Item                 | FP Before | FP After  |
|----------------------------|-----------|-----------|
| active neoplasia           | 18        | 4         |
| arrhythmia                 | 12        | 5         |
| acute coronary syndrome    | 9         | 2         |
| presence of dyspnea        | 7         | 3         |
| <b>Total (gated items)</b> | <b>46</b> | <b>14</b> |

Table 5: FP-gate effectiveness on dev80 EN (Mistral Large 3 teacher). The gates reduce FP for high-risk items by ~70% while preserving TP.

FP sources are: *chest pain* (13 FP), *cardiovascular diseases* (13 FP — triggered by the derivation rule from therapy flags), *level of consciousness* (9 FP — AVPU mapping errors), and *heart failure* (8 FP — “PMH: HF” in Stage 1 often produces FP when GT is unknown).

Comparing EN and IT dev80 error profiles: both languages share the same FN ceiling on rare conditions (*foreign body*, *agitation*), confirming that these are Stage 1 recall limitations independent of language. However, the *chest pain* FP problem is more severe in IT (13 FP vs. EN), likely because Italian notes use the term *dolore toracico* in broader

clinical contexts.

**Remaining FN errors.** The dominant error mode is **Stage 1 FN**: items not mentioned in the Stage 1 summary cannot be recovered by the deterministic Stage 2. High-FN items fall into three categories:

- **Rare conditions:** *foreign body in the airways* (11 FN in IT), *cardiac tamponade*, *aortic dissection* — these appear in <5% of notes and are almost never surfaced by Stage 1.
- **Implicit mentions:** *ab ingestis pneumonia*, *thoracic ultrasound* (6 FN in IT), *compression ultrasound (CUS)* — require domain-specific recognition that the SGR cascade may miss during condensation.
- **Context-dependent items:** *presence of dyspnea* (7 FN + 8 FP in IT), *level of consciousness* (6 FN + 9 FP in IT) — require nuanced temporal or categorical reasoning that is difficult to encode in a flat Key:Value format.

## 5. Discussion

### 5.1. Effectiveness of the SGR Decomposition

Applying SGR shifts the burden of structural correctness from prompt engineering to code. By having the LLM emit a stable, 9-key intermediate tree rather than directly producing 134 structured fields, we achieve **100% JSON parsing stability** even on quantized 4B edge models without grammar-constrained decoding. This is notable because

direct 134-field extraction from the same 4B model fails to produce valid JSON in >30% of cases in our preliminary tests.

The SGR decomposition also provides a natural **debugging boundary**: when the system produces an incorrect prediction, we can immediately determine whether the error originates in Stage 1 (the fact was not extracted) or Stage 2 (the fact was extracted but incorrectly mapped). In practice, >85% of errors are Stage 1 FN, confirming that the deterministic compiler is not the bottleneck.

## 5.2. Deterministic vs. LLM-Based Stage 2

Our experiments consistently show that the deterministic Stage 2 compiler outperforms LLM-based Stage 2 extraction in this task setting. We tested an LLM-based Stage 2 variant (using Anthropic Claude to directly extract 134 items from Stage 1 summaries) and found that it:

- introduces vocabulary drift (e.g., emitting `present` instead of `y`, `yes` instead of `certainly chronic`);
- produces inconsistent FP patterns across runs (non-deterministic);
- requires additional post-processing to fix format compliance;
- incurs non-trivial inference cost (\$0.01–0.03 per document via API).

The deterministic compiler, by contrast, achieves perfect vocabulary compliance, is fully explainable (every fill traces back to a Stage 1 line plus a specific rule), has zero inference cost, and guarantees bit-for-bit reproducibility.

## 5.3. SGR for Edge AI

This methodology is particularly enabling for **edge (local) deployments**. Large state-of-the-art models can handle 134 items in a single pass, but quantized 4B–8B local models lose track of the output schema mid-generation. Our SGR-constrained condensation fits within the reasoning bounds of smaller models (the 9-key JSON is ~400 tokens), securely delegating the complex vocabulary checks and item-level compilation to deterministic Python code. This makes the approach practical for privacy-preserving clinical deployments where sending patient data to external APIs is not acceptable. Concurrent work by [Ferrazzi et al. \(2026a\)](#) confirms that small LLMs (~1B parameters) can match or exceed larger baselines on Italian medical NLP tasks—including CRF filling—when combined with appropriate adaptation strategies such as few-shot prompting and constraint decoding.

Concretely, we serve MedGemma 1.5-4B-it as a Q5\_K\_M GGUF via `llama.cpp` on Apple M3 Pro (18 GB RAM). Stage 1 inference takes ~8 seconds per document; Stage 2 compilation is instant

(<10 ms). The entire dev80 evaluation completes in under 12 minutes on a single laptop.

## 5.4. Cross-Language Applicability

The pipeline naturally extends to Italian (`_it`) documents: the Stage 1 prompt is language-agnostic (it asks the model to extract clinical facts regardless of language), and Stage 2’s deterministic canonicalization handles multilingual item-name variants through the UMLS alias map. We prepared Italian test submissions using the same pipeline with no language-specific modifications beyond the `rules_v3e` Stage 2 variant (which adds tighter FP-drop gates for Italian-specific false positive patterns on *chest pain*, *heart failure*, and *arrhythmia*).

Cross-language results consistently favor Italian across all evaluation scales: dev10 (Table 4): IT F1=0.90 vs. EN F1=0.81; dev80 (Table 2): IT F1=0.6905 vs. EN F1=0.6543. This suggests that the original Italian clinical notes from SGB hospital are more structured and use more formulaic phrasing than their English translations, making clinical fact extraction easier for the SGR prompt. Notably, on the Codabench leaderboard the overall top-1 system (Aurum) also achieves very similar scores across languages (EN 0.68, IT 0.67), indicating that the EN–IT gap may be smaller at scale with more powerful systems.

## 5.5. Limitations of the 9-Key Condensation

The choice of exactly 9 domain keys represents a trade-off. Fewer keys (e.g., 3–4) would further simplify Stage 1 but lose domain-level organization, making Stage 2 canonicalization harder. More keys (e.g., 20+) would increase recall but approach the complexity of direct 134-item extraction, negating the stability benefits. We experimented with 5, 9, and 15 keys (“profile” parameter) and found that 9 keys offered the best balance of Stage 1 stability and Stage 2 recall.

## 6. Related Work

Clinical information extraction has a long history, from rule-based systems to modern LLM approaches. The i2b2/VA challenges ([Uzuner et al., 2011](#)) and ShARe/CLEF eHealth tasks ([Suominen et al., 2013](#)) established key clinical NLP benchmarks for structured information extraction from unstructured clinical text. More recently, large language models have demonstrated the potential to encode substantial clinical knowledge ([Singhal et al., 2023](#)), with domain-specific models like MedGemini ([Saab et al., 2024](#)) achieving strong results on medical question answering and clinical reasoning tasks.

**Structured Generation.** Several frameworks address the output stability problem in LLM-based extraction. Constrained decoding (Willard and Louf, 2023) enforces grammar constraints during token generation. SGLang (Zheng et al., 2024) provides programmatic control over LLM execution with RadixAttention for efficient structured output. XGrammar (Dong et al., 2024) offers a flexible engine for grammar-constrained generation. Our approach is complementary: rather than constraining the LLM’s token-level output, we simplify the target schema (9 keys vs. 134) and defer structural compliance to a deterministic post-processor. This avoids the computational overhead and backend-dependency of constrained decoding while achieving comparable output stability.

**Compiler-Based LLM Pipelines.** DSPy (Khat-tab et al., 2023) introduced the concept of compiling declarative language model programs into optimized pipelines. Our system shares the “compilation” philosophy: Stage 1 can be viewed as a declarative extraction step, and Stage 2 as a compiler that transforms the intermediate representation into the target format. However, our Stage 2 is fully hand-coded (deterministic), avoiding the need for optimization data or meta-learning.

**Ontology Grounding.** UMLS-based concept normalization via MetaMap (Aronson, 2001), QuickUMLS (Soldaini and Goharian, 2016), and ScispaCy (Neumann et al., 2019) has been widely applied for biomedical concept recognition.

Our work differs in its explicit two-stage decomposition with a fully deterministic Stage 2, the use of SGR patterns for intermediate representation stability, and the emphasis on edge-deployable privacy preservation. Concurrently, Ferrazzi et al. (2026a) present a systematic comparison of adaptation strategies for small LLMs on Italian medical NLP, covering 20 tasks including CRF filling, and demonstrating that fine-tuned 1.7B models can outperform 32B baselines.

## 7. Conclusion

LLM StructCore emphasizes contract alignment, format stability, and reproducible post-processing for CRF filling under extreme sparsity. The two-stage SGR decomposition isolates clinical extraction (Stage 1, any LLM) from strict submission enforcement (Stage 2, deterministic), enabling rapid iteration on Stage 1 recall and controlled-vocabulary grounding without sacrificing deterministic output guarantees. Our best submitted system achieves  $F1 = 0.63$  on the Codabench test set (EN), placing within 0.05 of the top-1 team (Aurum,  $F1 = 0.68$ ). The pipeline is fully language-

agnostic: Italian dev80 achieves  $F1 = 0.6905$  with no language-specific Stage 1 modifications, and the Italian test200 submission was prepared (but not submitted due to exhausted attempt limits). This demonstrates that the combination of SGR condensation and deterministic compilation is competitive across languages while offering full reproducibility and edge-deployment readiness.

Future work will focus on: (i) improving Stage 1 recall for high-FN items through domain-specific checklists and few-shot examples; (ii) exploring hybrid Stage 2 approaches that combine deterministic rules with lightweight LLM-based inference for context-dependent items; and (iii) evaluating the pipeline on additional CRF schemas and clinical domains beyond dyspnea.

## 8. Limitations

The deterministic Stage 2 **cannot recover facts missing from Stage 1 summaries**; thus, overall recall is Stage 1-limited. Some CRF items require nuanced contextual interpretation (e.g., distinguishing “past” vs. “current” respiratory distress) that may be lost in the 9-key condensation, increasing FN risk. The high FP cost of the task requires conservative evidence gating, which trades off recall for precision in borderline cases.

The abbreviation-based canonicalization relies on a manually curated map of 70+ abbreviations. New abbreviations or unconventional spellings not in the map will fail to match. The UMLS alias map mitigates this but is also static and may not cover all model-generated variants.

The FP-gate regex patterns are English-centric; while the Italian notes in this dataset are clinically structured and mostly use international terminology, extending the gates to fully cover Italian medical vocabulary would require additional curation.

Finally, our evaluation is limited to the official CL4Health 2026 data; generalization to other CRF schemas, clinical domains, or languages beyond English and Italian has not been tested.

## 9. Reproducibility

All deterministic logic is contained in the open-source package `llm_structcore`. Code will be provided in supplementary materials and open-sourced upon paper acceptance (Anonymous Repository). The repository includes:

- The `llm_structcore` Python package (Stage 1 prompt templates, Stage 2 compiler, scoring, submission builder).
- Scripts for running end-to-end inference across multiple backends (`llama.cpp`, OpenAI-compatible APIs, HF Transformers, NVIDIA API, Gemini Vertex, Anthropic).



- The complete UMLS alias map (data/umls\_crf\_mapping.json, 134/134 coverage).
  - The CRF ontology definition (data/ontology\_crf\_cl4health2026.md).
  - Organizer-provided evaluation scripts (external/CRF-filling-CL4Health2026/).
  - This system description paper (LaTeX source).
- The Stage 1 SGR schema guarantees json\_parse\_ok=100% with MedGemma 1.5-4B without requiring JSON-mode sampling. Stage 2 is fully reproducible: the same Stage 1 summary always produces the same 134-item submission, regardless of platform, Python version, or execution order.

## 10. Ethical Considerations

We use publicly available datasets under the organizer terms. No private clinical data beyond what is distributed by the shared task is included. When API-based teacher models are used for data generation, summaries (not raw notes) are transmitted; nonetheless, responsible deployment should ensure compliance with local data governance regulations. Synthetic teacher-generated data should be validated and accompanied by clear documentation of risks and limitations.

## 11. Bibliographical References

- Rinat Abdullin. 2026. Schema-guided reasoning (SGR): A complete guide. <https://abdullin.com/schema-guided-reasoning/>.
- Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proceedings of the AMIA Symposium*, pages 17–21.
- Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. 2024. XGrammar: Flexible and efficient structured generation engine for large language models. *Proceedings of Machine Learning and Systems*, 7.
- Pietro Ferrazzi, Mattia Franzin, Alberto Lavelli, and Bernardo Magnini. 2026a. Small LLMs for medical NLP: a systematic analysis of few-shot, constraint decoding, fine-tuning and continual pre-training in Italian. *arXiv preprint arXiv:2602.17475*.
- Pietro Ferrazzi, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. 2026b. Overview of the CRF 2026 shared task on clinical case report forms filling. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma, Mallorca (Spain). ELRA.
- Pietro Ferrazzi, Alberto Lavelli, and Bernardo Magnini. 2025. [Converting annotated clinical cases into structured case report forms](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 307–318, Vienna, Austria. Association for Computational Linguistics.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Mober, et al. 2023. DSPy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.
- 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*.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, et al. 2024. Capabilities of Gemini models in medicine. *arXiv preprint arXiv:2404.18416*.
- 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.
- Luca Soldaini and Nazli Goharian. 2016. Quick-UMLS: A fast, unsupervised approach for medical concept extraction. In *MedIR Workshop at SIGIR*.
- Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J.F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. 2013. Overview of the ShARe/CLEF eHealth evaluation lab 2013. In *CLEF*.
- Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. volume 18, pages 552–556.
- Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.



Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Jeff Huang, Chuyue Sun, Cody Hao Yu, Shiyi Cao, Christos Kober, Joseph E. Gonzalez, Clark Barrett, Ying Sheng, and Ion Stoica. 2024. SGLang: Efficient execution of structured language model programs. *arXiv preprint arXiv:2312.07104*.

## 12. Language Resource References

Anthony McEnery and others. 2004. *The EMILLE/CIIL Corpus*. EMILLE (Enabling Minority Language Engineering) Project. distributed via ELRA: ELRA-Id W0037, ISLRN [039-846-040-604-0](#).

Khalid Choukri and Niklas Paullson. 2004. *The Orientel Moroccan MCA (Modern Colloquial Arabic) database*. distributed via ELRA: ELRA-Id ELRA-S0183, ISLRN [613-578-868-832-2](#).

Roventini, Adriana and Marinelli, Rita and Bertagna, Francesca. 2016. *ItalWordNet v.2*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics “A. Zampolli”, National Research Council, in Pisa, ISLRN [532-206-426-067-2](#). PID <http://hdl.handle.net/20.500.11752/ILC-62>. Note: You don’t really need both an ISLRN and another PID, but it can’t hurt.

Speecon Consortium. 2011. *Catalan Speecon database*. SpeeCon. Speecon Project, distributed via ELRA: ELRA-Id S0327, Speecon resources, 1.0, ISLRN [935-211-147-357-5](#).

# GREYC at CRF Filling 2026: Rewrite Before You Extract - Rewriting Clinical Notes for Automated CRF

Jesus Lovon-Melgarejo, Jérémie Pantin, Gaël Dias

Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ,  
GREYC UMR6072, F-14000 Caen, France  
{jesus.lovon,jeremie.pantin, gael.dias}@unicaen.fr

## Abstract

This paper describes the system we submitted to the CRF-filling 2026 shared task. We propose a modular, LLM-based framework including an LLM as rewriter, which enhances the original clinical note from the perspective of each target CRF item; an LLM as an extractor, which retrieves the relevant value using a  $k$ -shot prompting strategy; and an LLM as a judge, which determines whether the clinical note contains evidence to support a given answer, defaulting to *unknown* otherwise. We evaluated our system on the English portion of the dataset; our complete framework achieves a macro-F1 of 0.64 on the development set. Our analysis reveals that while the rewriting step effectively generates correct factual information, it also increases false positives. The judge component mitigates this by adopting a conservative prediction strategy that substantially reduces false positives at the cost of a moderate reduction in true positives, yielding higher precision and better alignment with the shared task metric. On the test set, a light version of our system ranked 21 out of 32 public submissions, achieving a macro-F1 of 0.45.

**Keywords:** rewrite, LLMs, case report form

## 1. Introduction

A Case Report Form (CRF) is a standardised document widely used in clinical research to collect patient data across diverse studies and healthcare environments (Bellary et al., 2014). Each CRF contains a fixed set of items to be populated with patient medical information. Through their standardised structure, CRFs enable consistent data collection, ensuring accuracy, reliability, and validity required for reproducible clinical findings.

However, CRFs are typically filled manually from clinical notes and electronic medical records, a process that is time-consuming and prone to inconsistencies. Automating the population of CRFs from clinical narratives would therefore be highly beneficial: it could accelerate clinical research, reduce the manual workload of healthcare professionals, and produce structured representations of patient information. Consequently, recent research has explored automated systems that populate CRFs using information extracted from clinical notes and electronic medical records (Mac Kenzie et al., 2016; Gutiérrez-Sacristán et al., 2024).

The CRF filling 2026 shared task (Ferrazzi et al., 2026) aims to advance the development of systems applicable to real-world clinical settings. The datasets comprise CRFs for patients presenting dyspnea, using data from an Emergency Department in Italian and English.

Large Language Models (LLMs) have demonstrated remarkable capabilities in language understanding and generation (Petroni et al., 2019), with strong performance on health-related tasks such as clinical prediction, medical knowledge understand-

ing, and question-answering (Xu et al., 2024; Chen et al., 2023; Wu et al., 2024). Based on previous work that demonstrated that rewriting clinical patient profiles improves downstream performance, whether by generating new knowledge with LLMs (Sui et al., 2024), refining existing information using their internal knowledge (Lovon-Melgarejo et al., 2026), or augmenting profiles with external sources (Xu et al., 2024), we propose a multi-stage LLM framework for automated CRF filling.

In this work, we propose a multi-stage LLM-based framework for automated CRF filling. Our system consists of three components. First, an *LLM as rewriter* reformulates and enriches the original clinical input. Second, an *LLM as an extractor* retrieves the expected CRF item values, following previous approaches for structured information extraction (Ferrazzi et al., 2025). Third, to mitigate hallucination, where an LLM produces an answer despite insufficient supporting evidence (Su et al., 2022; Zhang et al., 2024), an *LLM as a judge* determines whether each extracted item can be reliably inferred from the provided context. Additionally, our approach accurately minimizes false positives and false negatives, preventing misleading entries.

## 2. Data

The CRF-filling dataset (Kaczmarek et al., 2026) consists of 290 anonymised emergency care clinical notes collected from the San Giovanni Bosco Hospital (Italy), covering patients admitted between January 2021 and December 2023. Each note is paired with a CRF capturing patient information

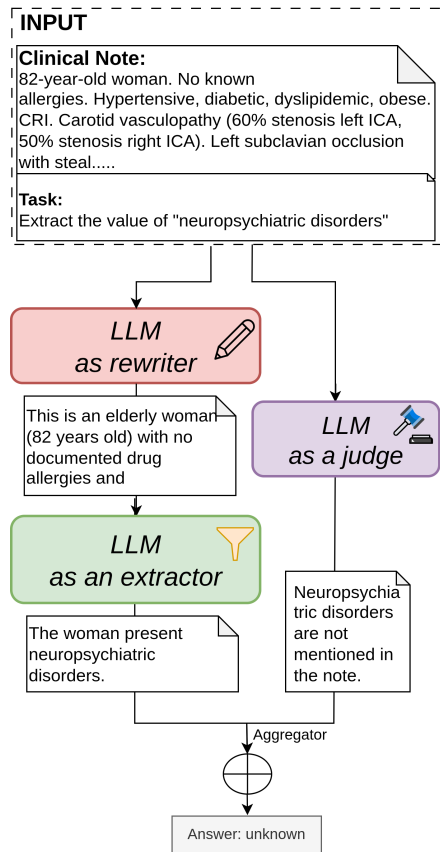


Figure 1: Overview of our framework architecture using three LLM-based components.

across a predefined set of medical items. The task focuses on the *dyspnea CRF*, which contains 134 items. The dataset is provided in both English and Italian and is split into training (10 notes), development (90 notes), and test (200 notes) sets.

Each CRF item is restricted to a predefined set of values depending on the type of information it represents, including binary (yes/no), ordinal (below/within/above), or measured/unknown categories used for tests without fixed categorical outcomes. The *unknown* value is valid for any item, enabling the schema to account for information absent from the clinical note. In addition to the manually annotated data, the task release includes 71 semi-automatically annotated note–CRF pairs in English (70 in Italian) (Ferrazzi et al., 2025), as well as 2667 unannotated clinical notes related to dyspnea, which were not used in our experiments.

### 3. Methodology

In this section, we describe our system, which comprises three LLM-based components: a rewriter, an information extractor, and a judge. We also consider an aggregator operator (Figure 1). Following previous work (Kaczmarek et al., 2026; Ferrazzi et al., 2025), we adopt a single-item processing

strategy to avoid cascading errors, processing each of the 134 CRF items independently, resulting in 134 inference steps per entry in the evaluation set. For each entry, our framework operates through two parallel branches. In the first branch, an *LLM as rewriter* explicitly enriches the clinical note before a second *LLM extracts* the relevant information for the target CRF item. In the second branch, an *LLM as a judge* evaluates whether the original input contains sufficient evidence to answer the item, acting as a safeguard against hallucinations introduced in the first branch.

#### 3.1. LLM as Rewriter

We prompt the LLM to complete the clinical note by interpreting and reorganising its content from the perspective of the target CRF item. This produces an enriched clinical note that either highlights information relevant to the item or explicitly states that no such information is present. Through this rewriting step, the parametric knowledge encoded in the LLM can be transferred and leveraged to enrich the clinical patient profile (Sui et al., 2024). The prompt used for the rewriter is as follows:

##### LLM prompt

You are an expert clinician specialized in clinical record analysis. Your task is to extract and interpret a specific clinical parameter from a patient’s clinical history. Your input is: a patient clinical history and a clinical question. Your output must be a structured clinical summary of the relevant facts from the profile, written so that a reader could easily answer the question.

PATIENT: {patient profile}

QUESTION: What is the link between the patient profile and {CRF item}?

#### 3.2. LLM as an Extractor

Following (Kaczmarek et al., 2026; Ferrazzi et al., 2025), we employ an LLM for information extraction using a *k*-shot prompting strategy, providing as input the concatenation of the original clinical note and the rewritten profile. The used prompt is:

#### LLM prompt

You are an expert clinical information extractor. You will be given a patient’s clinical history, an interpretation and a question. Your role is to answer the question strictly based on the provided information.  
 PATIENT: {patient profile}  
 INTERPRETATION: {rewritten patient profile}  
 QUESTION: What are the results and measures of {CRF item}?

### 3.3. LLM as a Judge

LLMs are prone to hallucination, generating plausible-sounding answers even when the clinical note provides no supporting evidence for a given CRF item. To mitigate this, we introduce an *LLM as a judge* tasked with determining whether a given clinical note contains sufficient information to reliably answer a target CRF item. The prompt used for this component is:

#### LLM prompt

You are an expert clinical information extractor. You will be given a patient’s clinical history, an interpretation and a question. Your role is to answer the question strictly based on the provided information.  
 PATIENT: {patient profile}  
 INTERPRETATION: {rewritten patient profile}  
 QUESTION: According to the clinical information provided, is the value of item explicitly mentioned or documented?

**Aggregator** Finally, the outputs of the *LLM as a judge* and *LLM as an extractor* components are aggregated to produce the final predicted value. Let  $J(n, q) \in \{\text{yes}, \text{no}\}$  denote the binary output of the *LLM as a judge*, indicating whether the answer to item  $q$  is present in clinical note  $n$ . Let  $E(n, q)$  denote the value predicted by the *LLM as an extractor* for item  $q$  given note  $n$ . The aggregator operator  $\oplus$  for an item  $q$  is defined as:

$$\oplus(n, q) = \begin{cases} E(n, q) & \text{if } J(n, q) = \text{'yes'} \\ \text{unknown} & \text{otherwise} \end{cases} \quad (1)$$

## 4. Experimental Setup

We evaluated our approach across multiple open weight LLMs. For general-purpose models, we

| Model             | Macro-F1                 |          |
|-------------------|--------------------------|----------|
|                   | w/o judge                | w/ judge |
|                   | w/o rewriter             |          |
| Med42             | 0.3441                   | 0.5691   |
| Llama             | 0.3776                   | 0.5846   |
| Qwen              | 0.3937                   | 0.5947   |
| Qwen <sub>s</sub> | 0.5046                   | 0.6293   |
|                   | w/ rewriter              |          |
|                   | Llama→Med42              | 0.3236   |
|                   | Qwen→Med42               | 0.3394   |
|                   | Qwen <sub>s</sub> →Med42 | 0.3716   |
|                   | Qwen <sub>s</sub> →Llama | 0.4275   |
|                   | Qwen→Qwen <sub>s</sub>   | 0.4496   |
|                   |                          | 0.5339   |
|                   |                          | 0.5883   |
|                   |                          | 0.6011   |
|                   |                          | 0.6026   |
|                   |                          | 0.6445   |

Table 1: Macro-F1 scores for the development (dev) set in English. We highlight **best** and second best results. Blue and red cells indicate our baselines and final framework, respectively.

tested Qwen<sub>s</sub><sup>1</sup>, Qwen<sup>2</sup>, and Llama<sup>3</sup>. We additionally considered one domain-specific model: MedLlama<sup>4</sup>(Med42), a medically-oriented LLM.

For the *extractor*, we applied  $k$ -shot prompting ( $k = 3$ ), selecting examples matching the target CRF item from the training split, and evaluated all four models. For the *rewriter*, we considered Qwen<sub>s</sub>, Qwen, and Llama. For the *judge*, we used the Qwen model.

Regarding inference parameters, the *rewriter* and *judge* used a temperature of 0.7, top-p of 0.9, and a maximum of 512 generated tokens. The *extractor* used a temperature of 0 and top-p of 1 to ensure deterministic outputs. For the *judge* specifically, rather than relying on generated output, we recover the logit values of the first valid generated token to determine a binary yes/no decision.

**Metrics** We evaluate our system using the official task metric, macro-F1. To enable a more in-depth analysis, we additionally report precision, recall, true positives, and false positives.

## 5. Results

We performed an in-depth evaluation on the development set, where gold annotations are available. Table 1 reports the results across all evaluated configurations. We report only the best configurations in our tables. Our baselines, using a 3-shot setup,

<sup>1</sup>Qwen/Qwen3-8B

<sup>2</sup>Qwen/Qwen3-32B

<sup>3</sup>meta-llama/Meta-Llama-3.1-70B-Instruct

<sup>4</sup>m42-health/Llama3-Med42-8B

| Model                    | w/o judge  |            |             |             | w/ judge          |                   |                     |                    |
|--------------------------|------------|------------|-------------|-------------|-------------------|-------------------|---------------------|--------------------|
|                          | TP↑        | FP↓        | P↑          | R↑          | TP↑(Δ%)           | FP↓(Δ%)           | P↑(Δ%)              | R↑(Δ%)             |
| <b>w/o rewriter</b>      |            |            |             |             |                   |                   |                     |                    |
| Med42                    | 264        | 2061       | 0.11        | 0.77        | 207(-22%)         | 174(-92%)         | 0.54(+391%)         | 0.58(-25%)         |
| Llama                    | 282        | 2078       | 0.12        | <b>0.87</b> | 231(-18%)         | 248(-88%)         | 0.48(+300%)         | 0.63(-28%)         |
| Qwen                     | 300        | 1522       | 0.16        | 0.85        | 211(-30%)         | <b>103</b> (-93%) | <b>0.67</b> (+319%) | 0.56(-34%)         |
| Qwen <sub>s</sub>        | 243        | <b>349</b> | <b>0.41</b> | 0.65        | 208(-14%)         | 107(-69%)         | 0.66(+61%)          | 0.55(-15%)         |
| <b>w/ rewriter</b>       |            |            |             |             |                   |                   |                     |                    |
| Llama→Med42              | 277        | 1910       | 0.13        | 0.83        | 203(-27%)         | 177(-91%)         | 0.53(+308%)         | 0.59(-29%)         |
| Qwen <sub>s</sub> →Med42 | 288        | 2040       | 0.12        | 0.83        | 222(-23%)         | 162(-92%)         | 0.58(+383%)         | 0.62(-25%)         |
| Qwen→Qwen <sub>s</sub>   | 277        | 1241       | 0.18        | 0.76        | 216(-22%)         | 159(-87%)         | 0.58(+222%)         | 0.58(-24%)         |
| Qwen→Med42               | 297        | 2060       | 0.13        | 0.86        | 229(-23%)         | 182(-91%)         | 0.56(+331%)         | <b>0.64</b> (-26%) |
| Qwen <sub>s</sub> →Llama | <b>309</b> | 1970       | 0.14        | 0.84        | <b>244</b> (-21%) | 170(-91%)         | 0.59(+321%)         | <b>0.64</b> (-24%) |

Table 2: Results from best configurations between wo/ judge (left) and w/ judge (right) configuration, and w/o rewriter (top) and w/ rewriter (top) configuration. We report the number of true positives (TP), false positives (FP), precision (P), and recall (R). **Green** cells indicate improvement with respect to the w/o judge configuration. Best results in **bold**.

achieve the highest performance with the Qwen model, reaching a macro-F1 of 0.5046. Particularly, general-domain LLMs consistently outperform the domain-specific model (MedLlama-Med42). Moreover, the setting without a rewriter (only extractor and judge) obtains the second-best results in terms of macro-F1 with a score of 0.6293, highlighting the impact of the *LLM as a judge* in this task.

Among the settings that include a rewriter but no judge, we observe a general trend of lower performance compared to the baselines, with a best macro-F1 of 0.4496 achieved by Qwen→Qwen<sub>s</sub> (where Qwen is the rewriter and Qwen<sub>s</sub> the extractor). When the full pipeline is used, our proposed system outperforms previous approaches, achieving a best macro-F1 of 0.6445 with Qwen<sub>s</sub>→Llama using *Qwen as a judge*. This substantial improvement suggests that jointly, the judge and rewriter components play a critical role.

The submitted system was based on a 3-shot setup and exhibited a trend consistent with development set findings, removing the *LLM as a judge* led to a drop in macro-F1 from 0.48 to 0.45.

### 5.1. Ablation Study

We further analyze the impact of the *LLM as a judge* and *LLM as rewriter* components on the development set performance. We rely on additional metrics for further examination of our approach. Table 2 compares the configurations with and without these LLM components across all models.

First, studying the effect of the rewriter, models with an *LLM as rewriter* achieve, on average, higher true positives (272.5 vs 289 w/o judge and 214.3 vs 222.8 w/ judge) and higher recall (0.79 vs 0.82, and

0.58 vs 0.61 for configuration w/o and w/ judge, respectively). These results suggest that rewriting effectively leverages LLM parametric knowledge to extract correct factual information. However, the expansion of the input context also leads to a substantial increase in false positives, i.e. cases where the correct value is *unknown* but the *LLM as an extractor* incorrectly assigns a value. This increase in false positives penalizes the macro-F1 score.

Second, analyzing the effect of *LLM as a judge*, our results show that applying this component consistently reduces true positives by at most 30%, yet achieves a reduction in false positives up to 92%, resulting in higher precision across all configurations. This suggests that *LLM as a judge* promotes a more conservative prediction strategy, improving the faithfulness of the extracted answers to the source text, which is a desirable behavior in sensitive domains such as healthcare.

## 6. Conclusion

In this paper, we described our system approach for the automated CRF filling from clinical notes, in the context of the CRF Filling 2026 shared task. Our framework combines three open-weight LLMs containing: an *LLM as rewriter*, an *LLM as an extractor*, and an *LLM as a judge*, the latter responsible for filtering out predictions unsupported by the clinical note. Our analysis showed that rewriting clinical notes prior to extraction effectively leverages the parametric knowledge of LLMs, recovering a higher number of correct factual values. However, this comes at the cost of increased false positives, as the expanded context encourages the model to assign values even when the answer should be



unknown. The *LLM as a judge* addresses this limitation by adopting a conservative prediction strategy, improving precision while reducing the true positives. Our complete framework achieves a score of 0.64 in terms of macro-F1 on the development set, confirming the complementary role of each component.

## 7. Acknowledgements

This work was supported by the PARTAGES project, winner of the Bpifrance France 2030 call for proposals “Digital Commons for Generative Artificial Intelligence”. The present work was performed using computing resources of CRIANN (Normandy, France).

## 8. Bibliographical References

- Shantala Bellary, Binny Krishnankutty, and MS Latha. 2014. Basics of case report form designing in clinical research. *Perspectives in clinical research*, 5(4):159–166.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.
- Pietro Ferrazzi, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. 2026. Overview of the CRF 2026 shared task on clinical case report forms filling. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma, Mallorca (Spain). ELRA.
- Pietro Ferrazzi, Alberto Lavelli, and Bernardo Magnini. 2025. [Converting annotated clinical cases into structured case report forms](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 307–318, Vienna, Austria. Association for Computational Linguistics.
- Alba Gutiérrez-Sacristán, Simran Makwana, Audrey Dionne, Simran Mahanta, Karla J Dyer, Faridis Serrano, Carmen Watrin, Pierre Pages, Sajad Mousavi, Anil Degala, et al. 2024. Development and validation of an open-source pipeline for automatic population of case report forms from electronic health records: a pediatric multi-center prospective study. *EBioMedicine*, 108.
- Gabriela Anna Kaczmarek, Pietro Ferrazzi, Lorenzo Porta, Vicky Rubini, and Bernardo Magnini. 2026. [Toward automatic filling of case report forms: A case study on data from an italian emergency department](#).
- Jesus Lovon-Melgarejo, Jose G. Moreno, Christine Damase-Michel, and Lynda Tamine. 2026. [Re-top: Learning to rewrite electronic health records for clinical prediction](#). WSDM ’26, page 458–468, New York, NY, USA. Association for Computing Machinery.
- William R Mac Kenzie, Arthur J Davidson, Andrew Wiesenthal, Jeffrey P Engel, Kathryn Turner, Laura Conn, Scott J Becker, Sharon Moffatt, Samuel L Groseclose, Jim Jellison, et al. 2016. The promise of electronic case reporting. *Public Health Reports*, 131(6):742–746.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473.
- Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. [Read before generate! faithful long form question answering with machine reading](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 744–756, Dublin, Ireland. Association for Computational Linguistics.
- 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.
- Zhenbang Wu, Anant Dadu, Mike Nalls, Faraz Faghri, and Jimeng Sun. 2024. Instruction tuning large language models to understand electronic health records. *Advances in Neural Information Processing Systems*, 37:54772–54786.
- Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May Dongmei Wang, Joyce Ho, and Carl Yang. 2024. [RAM-EHR: Retrieval augmentation meets clinical predictions on electronic health records](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 754–765, Bangkok, Thailand. Association for Computational Linguistics.
- Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng

Ji, and Tong Zhang. 2024. [R-tuning: Instructing large language models to say 'I don't know'](#). 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 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

# Innov8rs at CRF Filling 2026: An Iterative Multi-LLM Ensemble Pipeline with Dynamic Few-Shot Retrieval and Data-Driven Precision Filtering

Samminga Sainath Rao<sup>1</sup>, Sumit Mishra<sup>2</sup>, Chanchal Suman<sup>3</sup>

<sup>1</sup>Dept. of Computer Science and Engineering, RGIPT, Amethi, India

<sup>2</sup>Dept. of Computer Science and Engineering, NIT Warangal, Warangal, India

21cs2018@rgipt.ac.in, sumit@nitw.ac.in, chanchal@nitw.ac.in

## Abstract

In this paper, we present the technical report on the CL4Health 2026 Shared Task on Case Report Form (CRF) filling for our team Innov8rs. The paper explains the complete development of our system for the CL4Health 2026 Shared Task. We describe every phase of our system – from initial catastrophic failures with small models producing over 4,800 false positives, through prompt engineering breakthroughs, to our final multi-LLM ensemble combining Gemini 2.5 Flash and Llama 3.3 70B with dynamic TF-IDF-based few-shot retrieval. The main contribution of this work is a data-driven precision filter that suppresses predictions for CRF items with historically high false-positive rates. This single intervention reduced false positives from 816 to 171 on the English development set, boosting macro-F1 from 0.541 to 0.703. We document the engineering challenges of multi-API-key rotation across 11 Google API keys and 2 Groq keys, the design of four distinct ensemble strategies, and the critical analysis of why development-calibrated filters suffered from distribution shift on test data (final test F1: 0.47).

**Keywords:** Large Language Model, Few-Shot Prompting, Clinical NLP, Information Extraction, Hallucination Suppression, Model Ensembling, Precision Filtering

## 1. Introduction

The automated extraction of structured clinical variables from unstructured patient notes is a cornerstone of modern medical informatics, enabling secondary use of health data for epidemiological research, quality improvement, and clinical trial screening. The CL4Health 2026 Shared Task on Case Report Form (CRF) filling presents a particularly challenging scenario: participants must predict the value of 134 discrete medical items – spanning binary conditions, categorical states, vital sign interpretations, and free-text laboratory measurements – from Italian and English emergency department clinical notes documenting patients presenting with dyspnea (Ferrazzi et al., 2026b; Kaczmarek et al., 2026; Ferrazzi et al., 2025, 2026a).

The main challenge of this task is information *absence*. In any individual clinical note, the vast majority of the 134 CRF items are never mentioned. The correct prediction for an unmentioned item is “unknown”. This creates an extreme class imbalance problem: the ground truth is dominated by “unknown” values, and the scoring metric (macro-averaged F1 across exact string matches) *heavily penalizes* false positives – predictions of a definite value when the ground truth is “unknown”.

Large Language Models (LLMs) (Lee et al., 2020), while remarkably capable at extracting explicitly stated facts, exhibit a persistent and dangerous bias: they *hallucinate affirmative answers* (Xu et al., 2024; Maynez et al., 2020). When prompted to determine whether a patient has epilepsy from

a note that never mentions epilepsy, LLMs consistently predict “n” (No) rather than “unknown”. This behavior, which we term the **completion bias**, is rooted in the instruction-tuning objective that rewards helpful, complete responses.

For example, in case of clinical information extraction, this bias is catastrophic. For the item *pregnancy*, the correct answer is “unknown” (the note says nothing about pregnancy). However, every small model we tested predicted “n” (No), inferring a negative from absence. For *SpO2*, the correct answer is “95%” (the actual value), but weaker models predicted “measured” (a valid but less informative option). For *heart rate*, the model must interpret the raw number 110 against clinical thresholds and predict “tachycardic” (HR > 100).

Our contributions include: (1) a multi-provider extraction framework supporting OpenAI, Google, Anthropic, and Ollama; (2) a Schema Registry implementing strict constrained decoding against 134 valid option sets; (3) dynamic few-shot example retrieval using TF-IDF cosine similarity over a 2,667-note silver corpus; (4) a multi-API-key rotation framework managing 11 keys for throughput maximization; (5) four distinct ensemble strategies across model outputs; and (6) a data-driven precision filter achieving our breakthrough result.

Rest of the paper describes task description and data normalization in section 2. In section 3, we describe system architecture. The precision filter is discussed in section 4, and ensemble strategies is discussed in section 5. In section 6, the analysis of results is discussed, and finally in section 7, we

conclude the paper.

## 2. Task Description and Data Normalization

Given a clinical note  $d$  from an emergency department visit, the system must output predictions for all  $N = 134$  predefined CRF items. Each item  $i$  has a fixed set of valid values  $V_i$  and a special value “unknown”.

### 2.1. Schema Analysis and Constrained Normalization

We built a **Schema Registry** by parsing the provided `valid_options_train.json`. The module classified items into five types based on their valid option sets:

- **Binary** (e.g., *epilepsy*, *pregnancy*): {“y”, “n”}. Highest false-positive risk because LLMs default to “n” for unmentioned items.
- **Categorical** (e.g., *heart rate*): {“tachycardic”, “bradycardic”, “normocardic”}. Requires medical reasoning from raw numbers.
- **Chronic conditions** (e.g., *chronic pulmonary disease*): {“certainly active”, “possibly active”, “resolved”}.
- **Free-text/numeric** (e.g., *SpO2*, *creatinine*): Any numeric string (“95%”, “1.2”). Must extract actual values, not placeholders.
- **Lab status**: Both “measured” and actual values valid.

The normalization layer performed case-insensitive matching against valid options. Predictions not matching any valid option were coerced to “unknown”, preventing format-based scoring penalties.

---

#### Algorithm 1 Dynamic Few-Shot Example Selection

---

- 1: **Offline:** Index silver standard notes  $S = \{s_1, \dots, s_M\}$  with TF-IDF (max 10,000 features, English stop words removed)
  - 2: Compute TF-IDF matrix  $T \in R^{M \times F}$
  - 3: **for each** test note  $d$  **do**
  - 4:    $q \leftarrow \text{TF-IDF}(d)$  ▷ Query vector
  - 5:    $\text{sim}_j \leftarrow \cos(q, T_j) \quad \forall j \in [1, M]$
  - 6:    $\text{top3} \leftarrow \text{argsort}(\text{sim})[-3 :]$
  - 7:   Inject  $\{s_j : j \in \text{top3}\}$  with their silver annotations as few-shot examples into prompt
- 

## 3. System Architecture

Our final system consisted of four sequential stages each developed iteratively through failure analysis.

At first, TF-IDF retrieval is done, where top-3 silver notes are used. After the retrieval, V3 system prompt and dynamic few shot example is done. Output is extracted through LLMs, schema normalization is done. Using the conservative strategy, the ensemble merging is done. Finally the precision filter is done.

### 3.1. Prompt Engineering: Three Iterations

Our prompts evolved through three major versions, each addressing specific failure modes.

**V1 (Baseline):** A minimal system prompt instructing the LLM to extract CRF values. No guidance on handling missing information. Result: massive hallucination (> 4,800 FP on 80 patients).

**V2 (Open World Assumption):** Added explicit instructions: “‘unknown’ is the default value. Use when the text does not mention the item at all. Do not assume ‘n’ just because an item is missing.” Result: FP dropped 80% (to 916) but remained too high.

**V3 (Targeted FP Suppression + Clinical Rules):** Our best prompt (69 lines) incorporated three innovations: (a) *Explicit error examples* for the top 9 hallucinated binary items (pregnancy, trauma, thoracentesis, homelessness, pacemaker, CPR, chest pain, pneumothorax), each with the instruction: “If not mentioned, output ‘unknown’, NOT ‘n’.” (b) *Lab value extraction rules* forbidding the generic “measured” placeholder: “For SpO2: Extract the actual percentage value (e.g., ‘95%’). Do not output ‘measured’.” (c) *Vital sign interpretation thresholds*: BP  $\geq 140/90 \rightarrow$  “hypertensive”, HR  $> 100 \rightarrow$  “tachycardic”, Temp  $> 38^\circ\text{C} \rightarrow$  “hyperthermic”, RR  $> 20 \rightarrow$  “tachypneic”.

### 3.2. Dynamic Few-Shot Retrieval

Static few-shot examples cannot represent the diversity of 2,667+ clinical presentations (Brown et al., 2020; Rubin et al., 2022). We implemented retrieval-augmented in-context learning for English: for **Italian**, the silver standard IDs did not align with the clinical notes corpus. We selected three static examples from the gold and synthetic training sets, prioritizing notes with the most diverse positive predictions.

### 3.3. Multi-API-Key Rotation

Processing 200 test notes with prompts exceeding 6,000 tokens triggered severe rate limiting. We engineered a `KeyRotator` class managing 11 Gemini (Team et al., 2023) and 2 Groq keys:

In practice, all 11 Gemini keys exhausted within 15 – 20 minutes, yielding ~50 test notes per cycle. When Groq’s `llama-3.3-70b-versatile`

**Algorithm 2** Multi-Key Rotation Strategy

---

```

1: Initialize clients $C = \{c_1, \dots, c_{11}\}$, exhausted
 set $E = \emptyset$
2: for each test note d do
3: while extraction not successful do
4: if $|E| = |C|$ then
5: Wait 60s, reset $E \leftarrow \emptyset$
6: $k \leftarrow$ next key $\notin E$
7: result $\leftarrow$ EXTRACT(c_k, d , prompt)
8: if HTTP 429 or quota error then
9: $E \leftarrow E \cup \{k\}$, CONTINUE
10: else
11: Save result, BREAK

```

---

hit its 100K TPD limit, the script parsed the suggested retry time from the error message (e.g., “try again in 38m15s”) and slept accordingly.

#### 4. The Precision Filter: Our Key Contribution

The 816 FP from Gemini 2.5 Flash were not randomly distributed. Computing per-item precision:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \quad (1)$$

revealed a stark bimodal distribution: **34 items** had precision  $\geq 0.33$  (reliably extracted), while **100 items** had precision  $< 0.33$  (dominated by false positives). Many items had precision of exactly 0.0 – *every single prediction was wrong*. For example:

- *Abdominal pain*: TP = 0, FP = 23, Precision = 0.0
- *Phlebotomy*: TP = 0, FP = 18, Precision = 0.0
- *Respiratory failure*: TP = 0, FP = 29, Precision = 0.0

These items represented conditions that the LLM *always* inferred from indirect contextual clues (e.g., predicting “respiratory failure” whenever a patient presented with dyspnea) but were almost never explicitly confirmed in the ground truth.

| Threshold $\tau$ | Items Kept | TP         | FP         | F1           |
|------------------|------------|------------|------------|--------------|
| 0.0 (no filter)  | 134        | 352        | 816        | 0.541        |
| 0.1              | 62         | 332        | 298        | 0.654        |
| 0.2              | 49         | 321        | 226        | 0.672        |
| <b>0.33</b>      | <b>34</b>  | <b>305</b> | <b>171</b> | <b>0.703</b> |
| 0.4              | 29         | 286        | 129        | 0.689        |
| 0.5              | 24         | 259        | 96         | 0.668        |

Table 1: English Dev F1 Across Precision Filter Thresholds.

The best threshold was found to be  $\tau = 0.33$  (Table 1). Using this threshold, the filter removed 645 false positives (reducing them from 816 to 171)

while losing only 47 true positives (from 352 to 305). As a result, the macro-F1 score increased from 0.541 to 0.703, which corresponds to a 30% relative improvement achieved through a single post-processing step.

A similar trend was observed on the Italian development set. Applying the same method improved the F1 score from 0.500 to 0.683, reducing 640 false positives while losing only 39 true positives.

Table 2 isolates the contribution of each component.

| Config.            | TP  | FP   | FN  | F1           |
|--------------------|-----|------|-----|--------------|
| <i>English Dev</i> |     |      |     |              |
| V1 + Flash Lite    | 317 | 4814 | 22  | 0.355        |
| + V2 Prompt        | 280 | 916  | 55  | 0.465        |
| + V3 Prompt        | 306 | 616  | 52  | 0.496        |
| + 2.5 Flash        | 352 | 816  | 17  | 0.541        |
| + Prec. Filter     | 305 | 171  | 83  | <b>0.703</b> |
| <i>Italian Dev</i> |     |      |     |              |
| V1 + Flash Lite    | 315 | 6006 | 13  | 0.302        |
| + V2 Prompt        | 261 | 1584 | 77  | 0.434        |
| + 2.5 Flash        | 318 | 802  | 14  | 0.500        |
| + Prec. Filter     | 279 | 162  | 108 | <b>0.683</b> |

Table 2: Cumulative Contribution of Each Component

The precision filter alone accounts for more F1 improvement (+0.162) than model upgrade (+0.045), prompt V3 (+0.031), and prompt V2 (+0.110) combined.

#### 5. Ensemble Strategies for Test Submission

For the test set, we had Gemini 2.5 Flash predictions for 49/200 notes and Llama 3.3 70B for 200/200. We designed four ensemble strategies:

**S1 – Consensus:** Keep a prediction only when both models give the same non-“unknown” answer. If they disagree, the output is set to “unknown”. This approach focuses on improving precision by requiring agreement between the two models.

**S2 – Groq-Only:** Use predictions only from Llama 3.3 70B. This acts as a baseline with full coverage.

**S3 – Gemini-First:** Use Gemini predictions when they are available (for 49 notes), and use Groq predictions for the remaining 151 notes. This strategy gives priority to the stronger model when possible.

**S4 – Conservative (Best Performer):** For notes where both models provide predictions, if Gemini outputs “unknown”, the final result is also set to “unknown”, even if Groq predicts something else. If both models give non-“unknown” predictions, Gemini’s prediction is used. For notes where Gemini



predictions are not available, Groq’s predictions are used as they are.

| Test Submission                 | Macro-F1    |
|---------------------------------|-------------|
| S2: Groq-Only                   | 0.46        |
| S3: Gemini-First Merge          | 0.45        |
| S4: Conservative Ensemble       | <b>0.47</b> |
| Pure Gemini (49/200 coverage)   | 0.47        |
| Conservative + Precision Filter | 0.46        |

Table 3: Test set scores across submission strategies.

The Gemini-First merge (S3) scored *lower* than Groq-Only (S2), suggesting that Gemini introduced false positives on the notes it covered. The Conservative strategy (S4) mitigated this by using Gemini primarily as a filter rather than a source.

## 6. Analysis and Discussion

The most significant finding is the gap between development (0.703) and test (0.47) performance. More counterintuitively, applying the dev-calibrated precision filter to test predictions *decreased* the score from 0.47 to 0.46.

This exposes a fundamental limitation of post-hoc data-driven filtering: the 100 items blacklisted (based on dev precision) may represent genuine clinical findings present in Test patients but absent from the 80 dev patients. With only 80 calibration samples, per-item precision estimates are inherently unstable. Items with 0 TP and 5 FP on the Dev set might have 3 TP and 2 FP on the Test set – but our filter would suppress all 3 true positives.

### 6.1. Model Scaling Effects

A notable observation: the worst models produced the *most* non-unknown predictions (~60/note for Flash Lite) while the best models produced the fewest (6–10/note for Gemini 2.5 Flash). This inverse relationship between extraction density and quality confirms that **conservatism is the dominant strategy** in sparse extraction tasks under macro-F1 evaluation.

### 6.2. Key Takeaways

From the results, it is noticed that scaling model in terms of parameters, but is not always sufficient. It is also seen that, the precision filter (+0.162 F1) contributed more than all prompt iterations combined (+0.141 F1). However, it requires representative calibration data. With < 100 calibration patients, item-level statistics do not generalize robustly.

## 7. Conclusion

The CRF-filling experiment demonstrated that modern LLMs exhibit a strong completion bias when extracting sparse clinical information, requiring significant engineering efforts to mitigate it. Starting from an initial baseline that produced 4,814 false positives, our iterative pipeline—incorporating prompt engineering, larger model variants, dynamic retrieval mechanisms, and statistical filtering—reduced development false positives by 96% and improved the English development macro F1 score from 0.355 to 0.703. However, the discrepancy between development performance (0.703) and test performance (0.47) indicates the importance of more distribution-robust post-processing strategies. Future research should investigate: (1) instruction fine-tuning approaches, such as DPO or RLHF, that explicitly encourage an “unknown” response preference to address completion bias at the model level; (2) item-level confidence calibration based on generation probabilities; and (3) cross-validated precision estimation using bootstrap sampling to derive more reliable and generalizable item-level filtering methods.

## 8. Bibliographical References

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Pietro Ferrazzi, Mattia Franzin, Alberto Lavelli, and Bernardo Magnini. 2026a. [Small LLMs for Medical NLP: A Systematic Analysis of Few-Shot, Constraint Decoding, Fine-Tuning and Continual Pre-Training in Italian](#).
- Pietro Ferrazzi, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. 2026b. Overview of the crf 2026 shared task on clinical case report forms filling. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma, Mallorca (Spain). ELRA.
- Pietro Ferrazzi, Alberto Lavelli, and Bernardo Magnini. 2025. [Converting Annotated Clinical Cases into Structured Case Report Forms](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 307–318, Viena, Austria. Association for Computational Linguistics.
- Gabriela Anna Kaczmarek, Pietro Ferrazzi, Lorenzo Porta, Vicky Rubini, and Bernardo

Magnini. 2026. [Toward Automatic Filling of Case Report Forms: A Case Study on Data from an Italian Emergency Department.](#)

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.

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.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 2655–2671.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

# Aurum at CRF Filling 2026: Modular DSPy Extractors with Qwen3-Max for Multilingual CRF Filling

Vinay Babu Ulli<sup>1</sup>, Jyoti Kumari<sup>2</sup>, Anindita Mondal<sup>3</sup>

<sup>1</sup>Oogwai Analytics, Bangalore, India

<sup>2</sup>Department of Linguistics, Banaras Hindu University, India

<sup>3</sup>Language Technologies Research Center, IIIT Hyderabad, India

ullivinaybabu@gmail.com<sup>1</sup>, jyoti.bhu.ac.in<sup>2</sup>, anindita.mondal@research.iiit.ac.in<sup>3</sup>

## Abstract

This paper describes the submission by Team Aurum to the CL4Health @ LREC 2026 Shared Task on Case Report Form (CRF) Filling from dyspnea patient clinical notes. Extracting 134 structured clinical fields using a single Large Language Model (LLM) call often leads to schema-following errors, hallucination, and poor attention over complex instructions. To address this, we propose a modular extraction pipeline built with DSPy, which decomposes the massive CRF schema into 14 specialized, manually designed domain extractors (e.g., Medical History, Lab Values, Acute Diagnoses). Using *Chain-of-Thought* reasoning and strict Pydantic-typed validation with the Qwen3-Max (Thinking) model, our pipeline achieved an official Codabench Test Macro-F1 score of 0.68 in English and 0.67 in Italian, securing the 1st place ranking overall in the shared task. Through rigorous manual error analysis and iterative prompt optimization, we demonstrate that establishing explicit instruction boundaries between “not mentioned” (`None`) and “explicitly absent” (`False`) is critical for clinical information extraction, and that architectural modularity yields higher performance gains than simply scaling model parameters.

**Keywords:** Case Report Form, Clinical Information Extraction, Large Language Models, DSPy, Multilingual NLP

## 1. Introduction

The extraction of structured clinical variables from unstructured electronic health records (EHRs) is a foundational task in medical informatics, enabling downstream applications such as cohort selection, predictive modeling, and clinical trial matching (Wu et al., 2023). To support this, robust datasets are essential (Johnson et al., 2016). Recently, significant efforts have targeted the conversion of annotated clinical cases into structured data (Ferrazzi et al., 2025), including the automatic filling of Case Report Forms (CRFs) directly from emergency department notes (Kaczmarek et al., 2026).

Building upon this foundation, the CL4Health 2026 CRF:Filling shared task (Ferrazzi et al., 2026b) challenges systems to automatically extract 134 fields across English and Italian clinical notes describing patients presenting with dyspnea.

While Large Language Models (LLMs) have demonstrated remarkable capabilities in clinical text understanding (Singhal et al., 2023; Thirunavukarasu et al., 2023), the primary challenge of this task lies in the massive output schema and the strict semantic distinction between an explicitly negated clinical finding and an unmentioned one. Initial monolithic approaches, attempting to extract all 134 fields in a single LLM prompt, proved brittle. Models exhibited a strong tendency to hallucinate schema structures (Ji et al., 2023) and over-reason, inferring diagnoses from symptoms rather than strictly extracting what was explicitly

stated by the clinician (Lyu et al., 2023).

To overcome these challenges, we developed a highly modular, language-agnostic pipeline powered by the DSPy framework (Khattab et al., 2023). We partitioned the 134 fields into 14 distinct extractors, utilizing Qwen3-Max (Thinking) (Bai et al., 2023) coupled with step-by-step reasoning. This paper details our pipeline architecture, prompt optimization methodology, results, and comprehensive error analysis.

To facilitate reproducibility and further research, the complete source code, DSPy module definitions, and inference scripts are made publicly available on github at <https://github.com/vinayulli/crf-lrec-sharedTask>.

## 2. Methodology

### 2.1. Modular Extractor Architecture

Instead of relying on a single monolithic LLM call, which has been shown to degrade attention over long contexts and complex instructions (Liu et al., 2024b), we decomposed the CRF into 14 specialized extractors. We manually designed these categories based on clinical domain expertise and the natural semantic groupings of the target variables. This approach isolates semantic domains, allowing the LLM to focus on specific sections of the clinical note with targeted instructions (Zhou et al., 2022).

Figure 1 illustrates the overall data flow and architecture of our proposed system. An unstructured

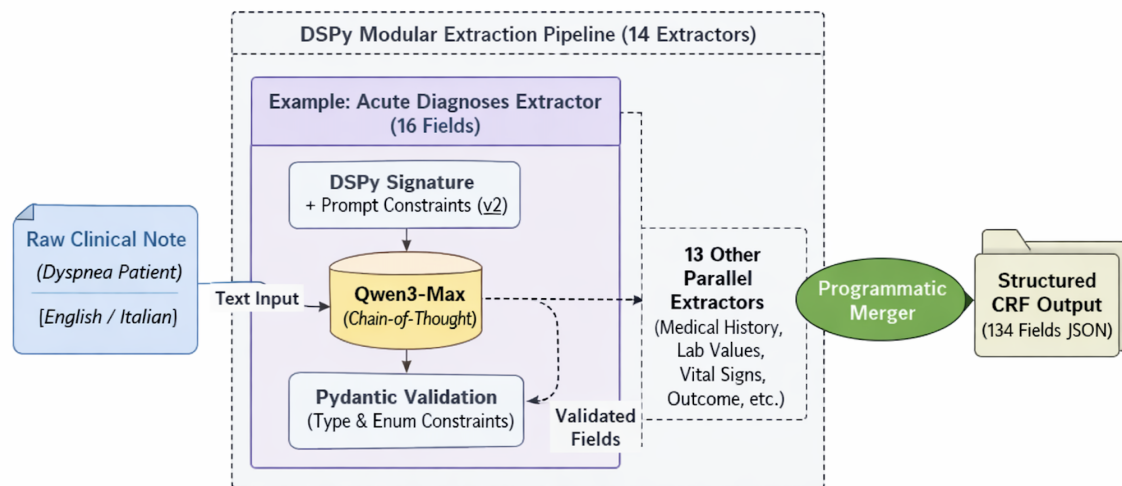


Figure 1: Architecture of the modular DSPy extraction pipeline, illustrating the parallel data flow from the raw clinical note to the final 134-field structured CRF.

clinical note is passed simultaneously to the 14 parallel extractors. Within each extractor module, the input is processed by a specialized DSPy signature combined with domain-specific prompt constraints. The Qwen3-Max model then applies step-by-step reasoning to extract the target variables, which are subsequently validated and type-cast using Pydantic schemas. Finally, a programmatic merger aggregates the validated fields from all 14 modules into the unified 134-field CRF JSON object.

The complete mapping of the 14 extractors to their corresponding 134 clinical features is detailed in Table 1.

Each extractor is implemented as a strongly-typed DSPy *Signature*. The output is strictly constrained using Pydantic (Colvin et al., 2023) sub-models, where all fields are typed as *Optional* and default to *None* (representing “unknown” or “not mentioned”).

## 2.2. Chain-of-Thought and Reasoning

For each signature, we utilized `dspy.ChainOfThought`. Before outputting the structured JSON, the model is prompted to generate a reasoning trace. This forces the model to sequentially locate explicit mentions in the text before assigning a boolean or categorical value, significantly reducing hallucination (Kojima et al., 2022). The 14 extractors run sequentially over each clinical note, and their outputs are programmatically merged into a unified 134-field object.

## 2.3. Prompt Engineering and Constraints

Initial experiments (Prompt v1) revealed a massive class of false positives (420 FPs in the dev

set). The root cause was the model’s inductive bias to interpret the absence of a mention as a definitive negative (*False*), a well-documented issue in LLM-based zero-shot extraction (Zheng et al., 2023).

To correct this, we implemented Prompt v2 via iterative manual refinement (Chen et al., 2023). We injected explicit instruction boundaries into the DSPy docstrings. For example, the v2 prompt for Acute Diagnoses explicitly stated:

*CRITICAL RULES: None means the condition is NOT MENTIONED AT ALL. This is the DEFAULT. False means the note EXPLICITLY STATES the condition is absent (e.g., “no pneumothorax”). Do not infer any diagnosis from symptoms alone.*

Furthermore, to ensure structured outputs, we constrained the generation using Pydantic *Enum* classes. For instance, categorical fields like *blood pressure* were restricted to exact string literals (*"normotensive"*, *"hypertensive"*, *"hypotensive"*). If the model’s *Chain-of-Thought* reasoning failed to map to these strict categories, the system safely fell back to *None*.

## 2.4. Zero-Shot Multilingual Transfer

Rather than employing translation steps or language-specific pipelines, which can introduce compounding translation errors (Conneau et al., 2020), we relied on the native multilingual alignment of modern foundation models (Devlin et al., 2019; Lin et al., 2022). This zero-shot transfer is particularly crucial given the specific terminological nuances and challenges of performing medical NLP natively in Italian (Ferrazzi et al., 2026a). The

| Extractor (# Fields)                  | Extracted Features                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|---------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Medical History</b> (22)           | chronic pulmonary disease, chronic respiratory failure, chronic cardiac failure, chronic renal failure, chronic metabolic failure, chronic rheumatologic disease, chronic dialysis, active neoplasia, cardiovascular diseases, diffuse vascular disease, dementia, neurodegenerative diseases, neuropsychiatric disorders, peripheral neuropathy, epilepsy/epileptic seizure, known history of epilepsy, history of alcohol abuse, history of drug abuse, history of allergy, history of recent trauma, immunosuppression, pregnancy |
| <b>Lab Values</b> (19)                | hemoglobin, leukocytes, platelets, creatinine, blood glucose, blood potassium, blood sodium, blood calcium, c-reactive protein, d-dimer, troponin, bnp or nt-pro-bnp, transaminases, serum creatinine kinase, inr, lactates, blood alcohol, blood drug dosage, urine drug test                                                                                                                                                                                                                                                       |
| <b>Acute Diagnoses</b> (16)           | pneumonia, ab ingestis pneumonia, pulmonary embolism, pneumothorax, acute coronary syndrome, heart failure, acute pulmonary edema, cardiac tamponade, aortic dissection, arrhythmia, severe anemia, intoxication, respiratory failure, asthma exacerbation, copd exacerbation, covid 19                                                                                                                                                                                                                                              |
| <b>Treatments</b> (14)                | administration of oxygen/ventilation, administration of bronchodilators, administration of diuretics, administration of fluids, administration of steroids, blood transfusions, cardio-pulmonary resuscitation, performance of thoracentesis, palliative care, antihypertensive therapy, anticoagulants or antiplatelet drug therapy, antiepileptic therapy already in place, poly-pharmacological therapy, compliance with antiepileptic therapy                                                                                    |
| <b>Imaging &amp; Diagnostics</b> (13) | chest rx, chest ct scan, abdomen ct scan, brain ct scan, brain mri, cardiac ultrasound, thoracic ultrasound, compression ultrasound (cus), ecg, ecg monitoring, eeg, pulmonary scintigraphy, gastroscopy (all evaluated for the presence of any abnormalities)                                                                                                                                                                                                                                                                       |
| <b>Epilepsy Assessment</b> (11)       | tonic-clonic seizures, further seizures in the ed, first episode of epilepsy, stiffness during the episode, eye deviation during the episode, pale skin during the episode, drooling during the episode, tongue bite, drowsiness confusion disorientation as postcritical state, duration of the patient's unconsciousness, duration of the patient's consciousness recovery                                                                                                                                                         |
| <b>Current Presentation</b> (10)      | presence of dyspnea, presence of respiratory distress, chest pain, agitation, blood in the stool, foreign body in the airways, general condition deterioration, head or other districts trauma, hemorrhage, concussive head trauma                                                                                                                                                                                                                                                                                                   |
| <b>Vital Signs</b> (7)                | blood pressure, heart rate, body temperature, respiratory rate, spo2, level of consciousness, level of autonomy (mobility)                                                                                                                                                                                                                                                                                                                                                                                                           |
| <b>Syncope Assessment</b> (7)         | situational syncope, tloc during effort, tloc while supine, supine-to-standing systolic blood pressure test, carotid sinus massage, presence of prodromal symptoms, situation description                                                                                                                                                                                                                                                                                                                                            |
| <b>Arterial Blood Gas</b> (4)         | ph, pao2, paco2, hco3-                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| <b>Social Context</b> (4)             | living alone, homelessness, need but absence of a caregiver, problematic family context                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| <b>Infection Screening</b> (3)        | influenza and various infections, sars-cov-2 swab test, neurologist consultation                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <b>Devices</b> (2)                    | presence of pacemaker, presence of defibrillator                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <b>Outcome</b> (2)                    | improvement of dyspnea, improvement of patient's conditions                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |

Table 1: Mapping of the 14 DSPy extractors to their corresponding 134 clinical features.

exact same English DSPy signatures and prompts were applied directly to the Italian clinical notes.

### 3. Experimental Setup

**Dataset:** Our pipeline was developed and evaluated using the official CL4Health dataset, consist-

ing of 80 Development documents and 200 Test documents per language (English and Italian) (Fer-razzi et al., 2026b). Evaluation on the test set was performed using the official Codabench Macro-F1 scorer (Xu et al., 2022). For our local development set evaluations, we calculated the Macro-F1 scores using the official evaluation script provided by the



organizers.<sup>1</sup>

**Models evaluated:** We evaluated 7 different foundation models during the development phase. These included Qwen3-Max (Thinking), Qwen3-8B (Bai et al., 2023), GPT-4o, GPT-4o-Mini (Achiam et al., 2023), Llama-4-Maverick (Touvron et al., 2023), and Gemma-3-12B-IT (Team et al., 2024) and Deepseek V3(Liu et al., 2024a).

## 4. Results

### 4.1. Development Set Evaluation

Table 2 reports the Macro-F1 scores obtained by various LLMs on the development set. We evaluated several multilingual foundation models, including Llama4 Maverik, GPT-4o, GPT-4o Mini, DeepSeek-V3, Gemma-3-12B-Instruct, and Qwen-series models. Among the evaluated models, Qwen3-Max (Thinking) with the optimized v2 prompts achieved the best performance on the English development set with a Macro-F1 score of 0.70, improving from 0.67 with the initial v1 prompt. GPT-4o achieved 0.68 on English and 0.65 on Italian, while GPT-4o mini produced comparable results (0.67 EN / 0.65 IT). The few-shot GPT-4o configuration slightly underperformed the zero-shot setup (0.65 EN / 0.63 IT). DeepSeek-V3 demonstrated strong multilingual performance, achieving 0.66 on English and 0.67 on Italian. Smaller models such as Qwen3-8B achieved competitive performance on English (0.67) despite having significantly fewer parameters.

| Model Name                  | Prompt    | English     | Italian     |
|-----------------------------|-----------|-------------|-------------|
| Llama-4-Maverick            | v2        | 0.58        | 0.58        |
| GPT-4o                      | v2        | 0.68        | 0.65        |
| GPT-4o few shot (3)         | v2        | 0.65        | 0.63        |
| GPT-4o-mini                 | v2        | 0.67        | 0.65        |
| Qwen3-Max (Thinking)        | v1        | 0.67        | 0.66        |
| <b>Qwen3-Max (Thinking)</b> | <b>v2</b> | <b>0.70</b> | <b>0.69</b> |
| Qwen3-8B                    | v2        | 0.67        | 0.60        |
| DeepSeek-V3                 | v2        | 0.66        | 0.67        |
| Gemma-3-12B-Instruct        | v2        | 0.62        | 0.60        |

Table 2: Development set Macro-F1 scores across different foundation models. The Qwen3-Max (Thinking) model with v2 prompts achieved the highest performance in both languages.

### 4.2. Official Test Set Results

Based on dev set performance, we selected Qwen3-Max (Thinking) with v2 prompts for our primary test submissions. As shown in Table 3, the

system demonstrated excellent language transfer, with only a marginal 0.01 F1 drop when applied zero-shot to the Italian test set.

| Language     | Macro-F1    |
|--------------|-------------|
| English (EN) | <b>0.68</b> |
| Italian (IT) | <b>0.67</b> |

Table 3: Official Test Set Results.

## 5. Error Analysis

To understand the limitations of LLMs in clinical CRF extraction, human annotators from our team conducted a rigorous manual error analysis on the development set predictions, identifying four primary root causes:

**1. “None” vs “False” Confusion (Over-prediction):** In v1, ~80% of all false positives stemmed from the model extracting `False` when a condition was simply not mentioned. While v2 prompts reduced FPs massively (from 420 down to 256), the model occasionally overcorrected, becoming too conservative and generating false negatives (increasing from 76 to 95) on genuinely absent conditions.

**2. Unwarranted Clinical Inference:** The *Chain-of-Thought* methodology occasionally harmed extraction by mimicking clinical diagnosis. For example, if an ECG showed an abnormality, the model inferred `arrhythmia = True` even if the clinician never explicitly documented the diagnosis. The strict ground truth standard clashes with the model’s natural diagnostic reasoning abilities.

**3. Lab Value Format Mismatches:** Because evaluation relies on exact string matching, variations in formatting severely penalized the system, a known limitation of lexical evaluation metrics in NLP (Kamalloo et al., 2023). For instance:

- GT: Hb 12 | Pred: 12 (Error: Missing prefix)
- GT: 24530 | Pred: WBC 24530 (Error: Added prefix)
- GT: 149 | Pred: 14.9 (Error: Unit conversion)

Instructing the model to “strip all labels” improved precision on some notes but generated errors on others where ground truth inconsistently retained labels.

## 6. Conclusion

our submission demonstrates that modularity is essential when extracting large, complex CRFs using LLMs. By breaking 134 fields into 14 distinct DSPy extractors and utilizing the reasoning capabilities

<sup>1</sup><https://github.com/hltfbk/CRF-filling-CL4Health2026>

of Qwen3-Max, we achieved highly competitive Macro-F1 scores of 0.68 (EN) and 0.67 (IT). Our analysis highlights that the true frontier in clinical information extraction is not merely increasing model size, but strictly bounding model reasoning to distinguish between implicit absence and explicit clinical negation, alongside overcoming the brittleness of exact-string matching evaluations.

## 7. Limitations

While our modular pipeline achieved state-of-the-art results, it introduces higher computational overhead and inference latency compared to a monolithic approach, as it requires 14 parallel LLM calls per clinical note. Additionally, the system's performance is bottlenecked by the exact-string matching evaluation metric, which occasionally penalizes clinically valid formatting variations (e.g., predicting "14.9 g/dL" instead of "149"). Finally, severe data sparsity in certain CRF domains (e.g., *Syncope* and *Outcome*) makes it difficult to reliably evaluate the model's true extraction capabilities for those specific fields.

## 8. Ethics Statement

This research utilizes de-identified clinical data provided by the CL4Health 2026 organizers, strictly preserving patient privacy. Because Large Language Models remain susceptible to hallucinations and incorrect medical inferences, our extraction pipeline is designed solely as an assistive tool for retrospective data structuring. It is not a substitute for human clinical judgment. Any real-world deployment of this system must incorporate robust human-in-the-loop validation to ensure patient safety and data integrity.

## 9. Acknowledgements

We thank the organizers of the CL4Health @ LREC 2026 shared task and Codabench for facilitating this challenge.

## 10. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge,

Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Samuel Colvin, Eric Jolibois, Hasan Ramezani, Adrian Garcia Badaracco, Terrence Dorsey, David Montague, Serge Matveenko, Marcelo Trylesinski, Sydney Runkle, David Hewitt, et al. 2023. Pydantic: Data validation using python type hints. In *pydantic: Data validation using Python type hints*.

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.

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.

Pietro Ferrazzi, Mattia Franzin, Alberto Lavelli, and Bernardo Magnini. 2026a. [Small llms for medical nlp: a systematic analysis of few-shot, constraint decoding, fine-tuning and continual pre-training in italian](#).

Pietro Ferrazzi, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. 2026b. Overview of the crf 2026 shared task on clinical case report forms filling. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma, Mallorca (Spain). ELRA.

Pietro Ferrazzi, Alberto Lavelli, and Bernardo Magnini. 2025. [Converting annotated clinical cases into structured case report forms](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 307–318, Viena, Austria. Association for Computational Linguistics.

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.

- Gabriela Anna Kaczmarek, Pietro Ferrazzi, Lorenzo Porta, Vicky Rubini, and Bernardo Magnini. 2026. [Toward automatic filling of case report forms: A case study on data from an italian emergency department](#).
- Ehsan Kamalloo, Nouha Dziri, Charles LA 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.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 9019–9052.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- 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.
- Qing Lyu, Josh Tan, Michael E Zapadka, Janardhana Ponnatapura, Chuang Niu, Kyle J Myers, Ge Wang, and Christopher T Whitlow. 2023. Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: results, limitations, and potential. *Visual Computing for Industry, Biomedicine, and Art*, 6(1):9.
- 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.
- 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*.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhakaran Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- 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).
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

## 11. Language Resource References

Alistair EW 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):1–9.

# sebis at CRF Filling 2026: A Two-Stage Local LLM Pipeline for Medical CRF Filling

Katharina Sommer<sup>†</sup>, Tristan Till<sup>†</sup>, Florian Matthes

Technical University of Munich, TUM School of Computation, Information and Technology  
{kathi.sommer, tristan.till, matthes}@tum.de

## Abstract

The extraction of structured clinical information from unstructured EHR notes is a persistent bottleneck in healthcare informatics. While large language models (LLMs) offer high performance, their deployment in clinical settings is hindered by privacy risks, inference costs, and the tendency to hallucinate beyond textual evidence. We address these challenges for the CL4Health 2026 Case Report Form (CRF) filling task by proposing a fully local, domain-adapted pipeline using the MedGemma-27B model. Our two-stage architecture, which separates binary presence classification from value extraction, enforces strict adherence to textual evidence and ensures deterministic outputs for negated, uncertain, or unknown states. By leveraging item-specific, few-shot in-context learning without external API calls or fine-tuning, our approach achieves a macro-F1 score of 0.55 on the official English test track. This result secures second place among all locally-hosted, open-source submissions. Our work demonstrates that privacy-preserving, on-premise LLM pipelines can achieve near-competitive performance with proprietary frontier models, providing a practical, data-sovereign framework for clinical NLP.

**Keywords:** Clinical NLP, Large Language Models, Case Report Forms, Few-Shot Learning, Local LLMs

## 1. Introduction

The extraction of structured clinical information from unstructured free-text notes remains a persistent bottleneck in healthcare informatics. Electronic health records (EHRs) contain rich patient narratives generated in high-volume settings such as emergency departments, yet converting these narratives into standardized, research-grade formats requires laborious manual review by clinicians or trained abstractors. Case Report Forms (CRFs) exemplify this challenge: they are predefined, study-specific instruments that demand consistent encoding of dozens of variables (e.g., comorbidities, vital signs, laboratory values, symptom chronicity) for each patient. The CRF Filling shared task (Fer-razzi et al., 2026) at the CL4Health 2026 workshop directly addresses this pain point by asking participants to automatically populate a CRF for dyspnea patients seen in emergency departments, using only clinical notes written in English or Italian. Success in the task would accelerate observational research, reduce documentation burden, and enable real-time, patient-friendly summarization.

While traditional rule-based and supervised clinical information extraction systems achieve high precision in narrow domains, they lack the scalability and zero-shot flexibility required for diverse clinical settings. Large language models accessed via commercial APIs offer impressive flexibility yet introduce critical drawbacks in real-world deployment, such as privacy and regulatory risks when sensitive patient data leave institutional firewalls,

prohibitive inference costs and latency at scale, and uncontrolled hallucinations or assumption-making that violate the strict “do not infer beyond the text” requirement of CRF filling.

Our submission closes these gaps through a fully local, domain-adapted LLM pipeline that requires only the official training set for few-shot guidance. By leveraging the open-weight MedGemma-27B model (Sjellerup et al., 2025), we guarantee data sovereignty and HIPAA/GDPR compliance while maintaining clinical-domain knowledge that generic models lack. The two-stage architecture, presence classification followed by value extraction, combined with engineered prompts and per-item few-shot exemplars, enforces strict adherence to the text and eliminates extraneous output. This design is both novel for the shared task and practically deployable in hospital environments. In summary, our core contributions are: (1) an applied demonstration of using a locally hosted medical LLM for CRF filling under strict privacy constraints; (2) a reusable, classification-then-extraction prompting framework that handles binary presence detection, categorical selection, and continuous “measured” values within a single inference pass; and (3) an efficient, zero-training implementation that achieves competitive performance in the English track while respecting the task’s emphasis on minimal supervision.

## 2. Related Work

Large Language Models (LLMs) have established few-shot prompting as a powerful paradigm for clin-

---

<sup>†</sup>These authors contributed equally to this work.



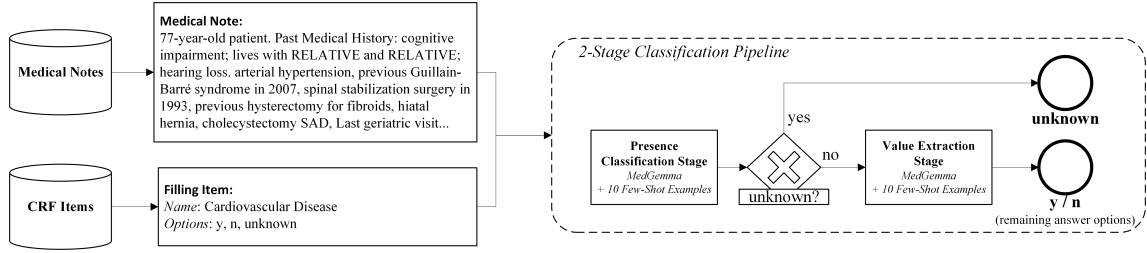


Figure 1: Overview of the two-stage classification-then-extraction pipeline. For each clinical note and CRF item, the system first determines presence via a classification prompt. If the presence is confirmed, the system proceeds to the Value Extraction Stage to output the specific categorical or measured value.

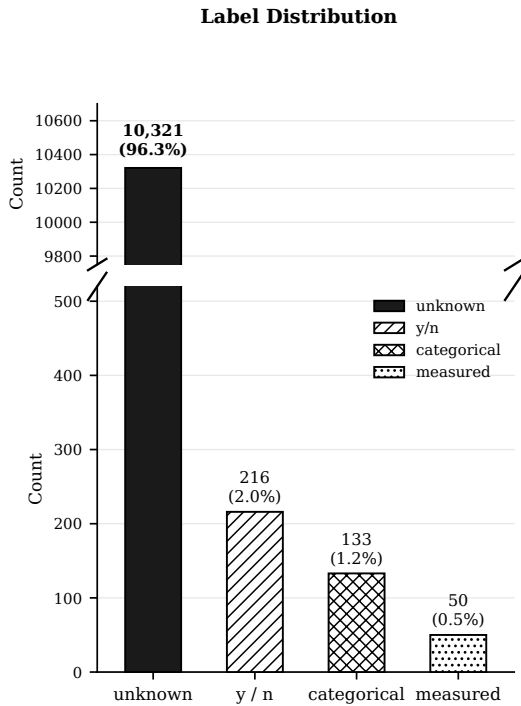


Figure 2: Label distribution of the training set across the four item types. The dataset exhibits a severe class imbalance, with "unknown" representing the vast majority of labels, emphasizing the challenge of minimizing false positives during extraction.

ical information extraction under low-resource constraints. Agrawal et al. (2022) demonstrated that models such as InstructGPT can achieve strong zero- and few-shot performance on clinical entity and relation extraction tasks directly from unstructured notes, relying solely on in-context examples without any fine-tuning. This approach aligns closely with the CL4health CRF-filling requirements, where systems must respect strict textual evidence, handle negations and uncertainty, and output deterministic values across multiple predefined items.

Further research has addressed the critical need for privacy-preserving deployment, focusing on fully local, open-source LLMs. Wiest et al. (2024) intro-

duced LLM-Alx, an on-premise pipeline based on Llama 2 that extracts structured information from clinical free text while keeping all patient data within institutional infrastructure. This line of work was extended by Bultjes et al. (2025), who showed that open-source generative LLMs deliver viable zero-shot performance on clinical extraction benchmarks in resource-constrained and multilingual settings, and by Richter-Pechanski et al. (2025), who achieved new state-of-the-art results on end-to-end medication information extraction through lightweight fine-tuning of local models on both English and German corpora.

Domain-specialized medical LLMs have further bridged the gap between general-purpose models and clinical precision. MedGemma-27B (Seller-gren et al., 2025), specialized on medical text, has established leading performance on clinical reasoning and extraction benchmarks while remaining fully open-weight and locally deployable. Our two-stage classification-then-extraction pipeline leverages exactly this model in a quantized, on-premise setup, synthesizing the few-shot flexibility shown by (Agrawal et al., 2022), the privacy guarantees of local open-source pipelines, and MedGemma’s domain knowledge into a practical solution for the CRF-filling shared task.

### 3. Methodology

#### 3.1. Data

We use the official English-track datasets released for the shared task (Kaczmarek et al., 2026; Ferrazzi et al., 2025). The training split contains clinical notes together with gold-standard annotations structured as a list of dictionaries. The test set provides the raw clinical notes and document identifiers required for submission. No synthetic data or external corpora were added, preserving the low-resource spirit of the task. Note that Figure 2 shows a severe class imbalance on a per-sample, per-CRF-item basis, in which "unknown" dominates.

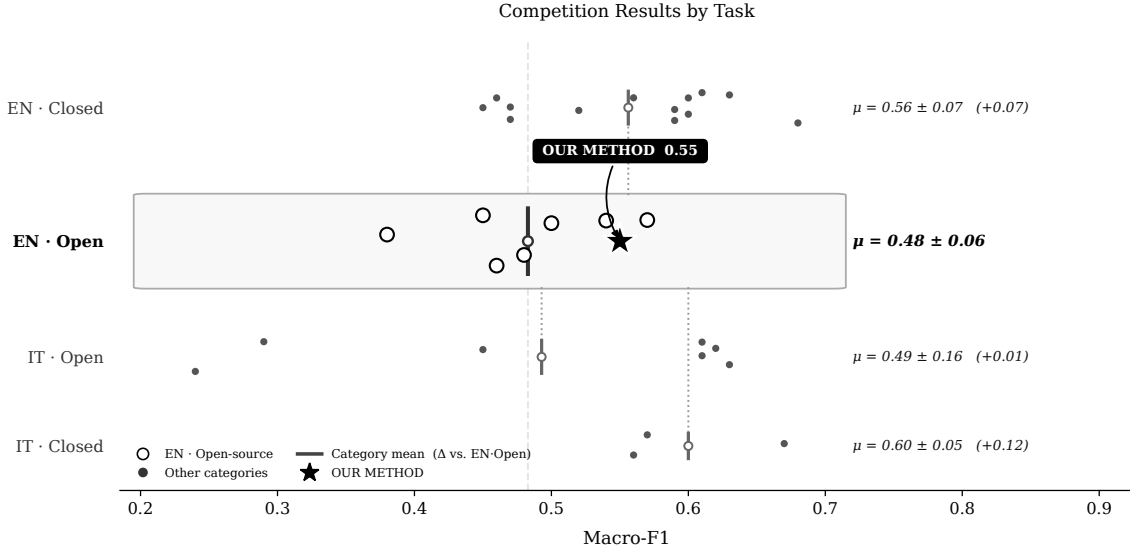


Figure 3: Diagram of the CL4Health 2026 CRF-filling leaderboard, split into detection language and open-vs. closed-weights models. On average, results on the Italian test set yield superior results to the English subset. Similarly, the use of closed-weights models shows higher Macro-F1 scores on average. Our method ranks #2 for submissions in English with open-weights models and #17 overall.

### 3.2. Model

Inference is performed with instruction-tuned MedGemma-27B (GGUF Q8\_K\_XL quantization) with temperature set to 0.

### 3.3. Pipeline Architecture

The system implements a two-stage, per-document, per-item loop, as shown in Figure 1. The algorithm for this approach can be seen in Algorithm 1.

**Presence Classification Stage:** We construct 10 few-shot examples for each category from the training. As the training set consists of only 10 samples, we use each sample to construct a few-shot example for every category. Then, the ground-truth labels are binarized to indicate whether the ground-truth is known (i.e., the answer is *"unknown"*) or an actual value. For *"unknown"*, we set the binary label to  $\bar{n}$  and for all other cases to  $\bar{y}$ . A classification prompt instructs the model to return exactly one character ( $\bar{y} / \bar{n}$ ) and includes explicit instructions on negations and uncertainty. If the model response is  $\bar{y}$ , we proceed with the Value Extraction Stage. Otherwise, *"unknown"* is recorded and the loop is aborted.

**Value Extraction Stage:** We construct another 10 few-shot examples using the actual ground-truth values, which now include the full option and or *"measured"* string. If the target item's option list begins with *"measured"*, we invoke a special

#### Algorithm 1 Two-Step LLM Extraction Pipeline (per document)

---

**Require:** Clinical note  $T$ , option table  $O$   
**Ensure:** Predictions  $P$

---

```

1: $P \leftarrow \emptyset$
2: for each row $r \in O$ do
3: $i \leftarrow r.item$
4: $opt \leftarrow r.options$
 $\triangleright$ Step 1: Presence classification
5: $fs_c \leftarrow create_few_shot(i)$
6: $p_c \leftarrow create_prompt_classification(i, T, fs_c)$
7: $\hat{y} \leftarrow LLM(p_c)$
8: if $\hat{y} = 'y'$ then
 $\triangleright$ Step 2: Value extraction
9: $fs_v \leftarrow create_few_shot(i, opt)$
10: if $opt = "measured"$ then
11: $p \leftarrow prompt_measured(i, T, fs_v)$
12: else
13: $p \leftarrow prompt_answer(i, opt, T, fs_v)$
14: end if
15: $\hat{v} \leftarrow LLM(p)$
16: else
17: $\hat{v} \leftarrow "unknown"$
18: end if
19: $P \leftarrow P \cup \{(i, \hat{v})\}$
20: end for
21: return P

```

---

measured-value prompt to task the model with extracting the specific value. Otherwise, a categorical prompt list is invoked, listing all and enforcing one of the allowed options.

| MedGemma | Few-Shot | Two-Step | Macro-F1 ( $\uparrow$ ) | %-gain ( $\uparrow$ ) | TPR ( $\uparrow$ ) | FPR ( $\downarrow$ ) | FNR ( $\downarrow$ ) |
|----------|----------|----------|-------------------------|-----------------------|--------------------|----------------------|----------------------|
| ✓        | X        | X        | 0.330                   | + 0.00%               | 7.35%              | 92.35%               | <b>0.30%</b>         |
| ✓        | ✓        | X        | <u>0.572</u>            | <u>+73.33%</u>        | <u>40.00%</u>      | <u>50.46%</u>        | 9.54%                |
| ✓        | X        | ✓        | 0.423                   | +28.18%               | 26.15%             | 66.00%               | <u>7.85%</u>         |
| ✓        | ✓        | ✓        | <b>0.604</b>            | <b>+83.03%</b>        | <b>40.56%</b>      | <b>44.66%</b>        | 14.79%               |

Table 1: Ablations on the internal development set quantifying the impact of the two-stage architecture and few-shot prompting on performance metrics (Macro-F1, TPR, FPR, and FNR). The baseline is a zero-shot MedGemma-27B model. Results demonstrate the additive benefits of the two-stage gatekeeping mechanism, reducing false-positive rates and boosting both recall and overall F1.

## 4. Results

### 4.1. Leaderboard Results

Our final submission achieves a macro-F1 score of 0.55 on the official hidden test set of the CL4Health 2026 CRF-filling shared task. As visualized in Figure 3, this result places us 17th out of 32 total submissions worldwide. Restricting the view to the English-only track, we rank 11th out of 22 participants. Most importantly, among the seven fully local and open-source systems submitted to the English track, our MedGemma-based pipeline secures a strong 2nd place, coming within 0.02 macro-F1 of the top local open-source entry (0.57) while remaining more than 0.13 points behind the overall winner (a closed-source English system at 0.68). These rankings highlight that our privacy-preserving, on-premise approach delivers highly competitive performance without relying on external APIs or proprietary models.

### 4.2. Development Phase Ablations

To justify each design decision, we performed systematic ablations on an internal development split derived from the official training set. Starting from a naive zero-shot MedGemma-27B baseline (macro-F1 = 0.330), simply adding the two-stage classification-then-extraction architecture increased performance to 0.423, confirming that first verifying presence before attempting value extraction substantially reduces false-positive predictions on absent items. Incorporating item-specific 10-shot examples without the two-stage design further raised the score to 0.572, demonstrating the critical value of in-context learning. Finally, the complete pipeline including two-stage prompting plus few-shot examples reached 0.604 macro-F1 (+0.032 over the few-shot-only variant), proving that all components act additively. These gains directly motivated our final architecture and explain why the same pipeline translated into the strong leaderboard ranking reported above.

## 5. Discussion

The leaderboard results (Figure 3) confirm a performance gap between closed-weight and open-weight models. While proprietary frontier models maintain advantages in scale and alignment, our second-place finish among local, open-source submissions demonstrates that domain-adapted models like MedGemma-27B can achieve near-competitive performance while maintaining data sovereignty: a requirement for clinical deployment.

Ablation studies (Table 1) quantify the additive benefits of our pipeline. A naive zero-shot baseline is overly conservative, missing most positive cases. Integrating few-shot examples provides the single largest gain, teaching the model to identify positive evidence. Separately, the two-stage gatekeeper improves precision by reducing spurious extractions. Combining both components yields the optimal error profile, confirming that few-shot guidance teaches detection while the classification stage acts as a filter against over-prediction.

Our primary development challenge was mitigating false positives, particularly the model’s tendency to over-interpret negated or distant mentions as positive evidence. Reliably distinguishing between a ruled-out label and missing information remains a known difficulty in clinical NLP. We addressed this through explicit negation/uncertainty instructions, the two-stage gatekeeper, and strict output constraints. Despite these improvements, residual false positives remain the primary bottleneck. Future work will explore lightweight preference-based alignment, retrieval-augmented few-shot selection, and ensemble-based uncertainty estimation to further suppress over-confident extractions without sacrificing recall.

## 6. Future Work

While our two-stage pipeline provides a privacy-preserving and efficient alternative to commercial APIs, several areas of future work remain. First, future work will explore automated retriever-based selection to dynamically choose the most relevant exemplars for each patient note.

Second, although our presence-classification gatekeeper reduces false positives, the model still occasionally misinterprets negated mentions as positive evidence. Future developments will explore lightweight, parameter-efficient fine-tuning (e.g., LoRA) on domain-specific negation-handling datasets to improve precision. Finally, our pipeline currently operates per-item; integrating a cross-item consistency checker could further reduce internal contradictions (e.g., ensuring mutually exclusive clinical symptoms are not both marked as present).

## 7. Conclusion

In this paper, we presented a privacy-preserving, two-stage pipeline for CRF filling using the MedGemma-27B model. By combining a presence-classification gatekeeper with value-extraction prompts, our system effectively mitigates the common LLM pitfalls of over-interpretation and hallucination. Our results confirm that local, open-weight models, when paired with strategic few-shot guidance, can bridge the performance gap between institutional, on-premise solutions and closed-source commercial APIs. This approach provides a scalable and secure alternative for clinical extraction tasks where data sovereignty is a non-negotiable requirement. Future work will focus on integrating preference-based alignment and ensemble-based uncertainty estimation to further refine the precision of clinical extractions in high-stakes healthcare environments.

## 8. Limitations

Experiments were conducted only on English data, limiting the scope to English-only health records and omitting comparisons of how the approach and model perform in non-English languages.

Furthermore, the dependency on few-shot in-context learning makes the system sensitive to the quality and diversity of training examples. Without at least a few gold standard examples, the accuracy of this approach cannot be guaranteed.

Lastly, the original dataset is in Italian, the dataset used for experiments is translated into English by the dataset authors. Therefore, translation artifacts could be present in the dataset, potentially skewing the results of the task.

## 9. Ethical Statement

This work uses anonymised patient data collected at San Giovanni Bosco Hospital. The data was anonymised by the dataset authors and made publicly available. Therefore, no ethical considerations apply.

## 10. Bibliographical References

- 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.
- Luc Builtjes, Joeran Bosma, Mathias Prokop, Bram van Ginneken, and Alessa Hering. 2025. Leveraging open-source large language models for clinical information extraction in resource-constrained settings. *JAMIA open*, 8(5):ooaf109.
- Pietro Ferrazzi, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. 2026. Overview of the crf 2026 shared task on clinical case report forms filling. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma, Mallorca (Spain). ELRA.
- Pietro Ferrazzi, Alberto Lavelli, and Bernardo Magnini. 2025. [Converting annotated clinical cases into structured case report forms](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 307–318, Viena, Austria. Association for Computational Linguistics.
- Gabriela Anna Kaczmarek, Pietro Ferrazzi, Lorenzo Porta, Vicky Rubini, and Bernardo Magnini. 2026. [Toward automatic filling of case report forms: A case study on data from an italian emergency department](#).
- Phillip Richter-Pechanski, Marvin Seiferling, Christina Kiriakou, Dominic M Schwab, Nicolas A Geis, Christoph Dieterich, and Anette Frank. 2025. Medication information extraction using local large language models. *Journal of biomedical informatics*, page 104898.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.
- Isabella Catharina Wiest, Fabian Wolf, Marie-Elisabeth Leßmann, Marko van Treeck, Dyke Ferber, Jiefu Zhu, Heiko Boehme, Keno K Bressen, Hannes Ulrich, Matthias P Ebert, et al. 2024. Llm-aix: An open source pipeline for information extraction from unstructured medical text based on privacy preserving large language models. *MedRxiv*.

## A. Appendix

### A.1. Prompts

The following prompt shows the first step of the pipeline, which classifies whether a given attribute appears in the clinical note. The model is expected to return either `y` or `n` depending on whether the given category is present in the clinical text or not.

#### Presence Classification Prompt (1<sup>st</sup> Step)

You are a clinical assistant helping a doctor extract information from clinical text.

For a given category, determine whether information about it is present in the text.

##### Instructions:

- Only return a single character: 'y' for Yes (present) or 'n' for No (absent).
- Do NOT make any assumptions beyond the text.
- Pay attention to negations, unknown, or unclear mentions.
- Do not add explanations or extra text.

Few-shot examples:  
{few\_shot}

Now check the following category in the given clinical note:

Category: {category}

Clinical Text: {text}

Answer:

The next two prompts show the prompts for the actual CRF task. There are two types of categories: one is selecting the correct value from a list of options (see the first prompt), and the other is extracting an exact numerical value with a unit from the text (see the second prompt). The model is prompted with either one of the options depending on the category.

#### Value Extraction Prompt (2<sup>nd</sup> Step)

You are a clinical assistant helping a doctor extract information from clinical text.

Your task is to select the correct value for a given category based ONLY on the text.

You MUST choose from the provided options and nothing else.

##### Instructions:

- Only return ONE of the provided options.
- The option 'unknown' means the clinical text does NOT provide enough information about this category.
- Do NOT make assumptions beyond the text.
- Pay attention to negations, uncertainty, or unclear mentions.
- Do not provide explanations or extra text; return only the single selected option.

Few-shot examples: {few\_shot}

Now evaluate the following:

Category: {category}  
Options: {option}  
Clinical Text: {text}

Answer:

#### Value Extraction Measured Prompt (2<sup>nd</sup> Step)

You are a clinical assistant helping a doctor extract information from clinical text.

Your task is to find the value of a given category from the text. You MUST choose from the text itself and include the unit if present.

##### Instructions:

- Extract the exact value from the clinical text, including units if available.
- If the information is not mentioned in the text, return 'unknown'.
- Do NOT make assumptions or invent values beyond what is in the text.
- Pay attention to negations, uncertainty, or unclear mentions.
- Do not provide explanations or extra text; return only the single extracted value or 'unknown'.

Few-shot examples: {few\_shot}

Now evaluate the following:

Category: {category}  
Clinical Text: {text}



Answer:

The following prompt shows how one few-shot example looks like for the classification task in the 1<sup>st</sup> Step. The answer is either `y` or `n`, depending on whether the category appears in the clinical text or not.

#### Presence Classification Few Shot Template (1<sup>st</sup> Step)

```

Category: {category}
Clinical Text: {note}
Answer: {value}

```

The next prompt shows one few-shot example for the actual CRF task. The answer is either one of the given options or the exact extracted value with a unit.

#### Value Extraction Few Shot Template (2<sup>nd</sup> Step)

```

Category: {category}
Options: {options}
Clinical Text: {note}
Answer: {value}

```

## A.2. Prompt Examples

This section gives an example prompt for every stage of the pipeline. Assume we want to fill in the category "chronic pulmonary disease" which has the following options: ['certainly chronic' 'possibly chronic' 'certainly not chronic' 'unknown']. The following examples shows how the prompt would look like. For better readability only 2 few-shot samples are shown.

#### 1<sup>st</sup> Step - Presence Classification

You are a clinical assistant helping a doctor extract information from clinical text.

For a given category, determine whether information about it is present in the text.

Instructions:

- Only return a single character: 'y' for Yes (present) or 'n' for No (absent).

- Do NOT make any assumptions beyond the text.
- Pay attention to negations, unknown, or unclear mentions.
- Do not add explanations or extra text.

Few-shot examples:

Category: chronic pulmonary disease  
Clinical Text: TRIAGE:  
Reports having fallen accidentally, impacting:  
    - occipital region with no loss of consciousness or concussion  
    - right trochanter  
    - right knee  
    - right ankle  
CS 15\_Cincinnati negative\_fluent speech\_isochoric, isocyclic, and photoreactive pupils. No rigidity. Conscious, lucid, oriented, not agitated.  
wearing rigid cervical collar  
BP 120/80 mm/hg, SpO2 99% on room air, RF 14 apm, HR 80 bpm r

MEDICAL ASSESSMENT:

    reports accidental fall to the ground at home following loss of support on right leg (spontaneous femur fracture?). Impact to the right side of the body, as described in triage. Currently experiencing pain in the right femur and right elbow. Head trauma without loss of consciousness. No vomiting after the episode.

    No other details at the moment

Past Medical History:

- Arterial hypertension
- CAD (reports quadruple BPAC)
- Diabetes mellitus type 2
- Left hip prosthesis

Therapy: metformin 500 mg, iperten 1/2 tablet, enalapril 5 mg, omnic 0.4 mg, seloken 100 mg 1/2 tablet x 2, cardioASA, torvast 20 mg, contramala 15 drops, tachipirina AD, omnepraozlo 20 mg

ALLERGIES: none known

Answer: n

-----

Category: chronic pulmonary disease  
 Clinical Text: Alert, oriented, cooperative  
 BP 150/80 mmHg HR 78r bpm SpO2 96% on room air RF 18 apm Temp 36.8  
 Pulmonary examination: Bilateral diffuse vesicular murmur, no additional sounds.  
 Cardiac examination: valid, rhythmic tones  
 Abdominale examination: distended, palpable, not tender or painful, Murphy and Blumberg negative, peristalsis present  
 Lower limbs: no peripheral edema; peripheral pulses hyposphygmia, bilaterally palpable up to the popliteal artery, with no pedidum pulse palpable bilaterally; dry necrosis left first toe, mild cyanosis in the remaining toes of the left foot; no sensory or motor deficits in the left lower limbs  
 ECG unchanged from the previous one

Answer: n

-----  
 {8 more few-shot samples - skipped here for readability}  
 -----

Now check the following category in the given clinical note:

Category: chronic pulmonary disease  
 Clinical Text: 77-year-old patient.  
 Past Medical History: cognitive impairment; lives with RELATIVE and RELATIVE; hearing loss. arterial hypertension, previous Guillain-Barr syndrome in 2007 spinal stabilization surgery in 1993, previous hysterectomy for fibroids, hiatal hernia, cholecystectomy  
 SAD, Last geriatric visit  
 PHONE\_NUMBER: MRI cerebral vasculopathy - presence of buccal dyskinesias - slow walking with risk of falls  
 Therapy: pantoprazole 40-olmesartan 1/2- xanax 0.25- contramal 4 drops x2, levopraid 1/2 x 2, folic acid, stilnox,  
 RELATIVE PHONE\_NUMBER

Recent Medical History: patient arrived last night for an episode of sudden illness characterized by a scream followed by generalized TC tremors in all 4 limbs; this morning a second fit was seen in the Emergency and Acceptance Department, TC in all 4 limbs, preceded by a scream.  
 Performed in the Emergency Department:  
 blood tests: no abnormalities (normal electrolytes), no fever; experiencing sleep loss due to frequent nighttime awakenings. Also performed brain CT brain without contrast which showed a picture of discrete diffuse cortico-subcortical atrophy, no acute manifestations; known chronic vascular disease.

Neurological Examination: patient alert, cooperative, rather quiet, partially oriented in time/place; no clear oculomotor deficits, blinks at bilateral visual threat, no underleveling at Mingazzini I; raises lower limbs alternately for 4-5 sec with symmetrical controlled descent. No appreciable sensory deficits. Bilateral plantar reflex in flexion.

Conclusions: generalized TC seizures (2 episodes) due to moderate cerebral atrophy and cognitive impairment.

Recommendation to start gardenale 50 mg 1 tablet in the evening. Clinical observation until tomorrow morning; if there are no other episodes or other illnesses, discharge tomorrow with indication to perform EEG with priority B and check-up in 6 months at epilepsy clinic with blood phenobarbital levels, a complete blood count, and liver and kidney function tests..

Answer:

Assuming that the answer of the LLM to the above prompt is y, and because the option is not "measured" but categorical, the following prompt is sent to the LLM. Again, only 2 of the few-shot prompts are shown.

## 2<sup>nd</sup> Step - Value Extraction

You are a clinical assistant helping a doctor extract information from clinical text.

Your task is to select the correct value for a given category based ONLY on the text.

You MUST choose from the provided options and nothing else.

### Instructions:

- Only return ONE of the provided options.
- The option 'unknown' means the clinical text does NOT provide enough information about this category.
- Do NOT make assumptions beyond the text.
- Pay attention to negations, uncertainty, or unclear mentions.
- Do not provide explanations or extra text; return only the single selected option.

### Few-shot examples:

Category: chronic pulmonary disease

Options: ['certainly chronic' 'possibly chronic' 'certainly not chronic' 'unknown']

Clinical Text: TRIAGE:

Reports having fallen accidentally, impacting:

- occipital region with no loss of consciousness or concussion
- right trochanter
- right knee
- right ankle

CS 15\_Cincinnati negative\_fluent speech\_isochoric, isocyclic, and photoreactive pupils. No rigidity. Conscious, lucid, oriented, not agitated.

wearing rigid cervical collar

BP 120/80 mm/hg, SpO2 99% on room air, RF 14 apm, HR 80 bpm r

### MEDICAL ASSESSMENT:

reports accidental fall to the ground at home following loss of support on right leg (spontaneous femur fracture?). Impact to the right side of the body, as described in triage. Currently experiencing pain in the right femur and right elbow. Head trauma without loss of consciousness. No vomiting after the episode. No other details at the moment

### Past Medical History:

- Arterial hypertension
- CAD (reports quadruple BPAC)
- Diabetes mellitus type 2
- Left hip prosthesis

Therapy: metformin 500 mg, iperten 1/2 tablet, enalapril 5 mg, omnic 0.4 mg, seloken 100 mg 1/2 tablet x 2, cardioASA, torvast 20 mg, contramala 15 drops, tachipirina AD, omnepraozlo 20 mg  
ALLERGIES: none known

Answer: unknown

Category: chronic pulmonary disease

Options: ['certainly chronic' 'possibly chronic' 'certainly not chronic' 'unknown']

Clinical Text: Alert, oriented, cooperative

BP 150/80 mmHg HR 78r bpm SpO2 96% on room air RF 18 apm Temp 36.8

Pulmonary examination: Bilateral diffuse vesicular murmur, no additional sounds.

Cardiac examination: valid, rhythmic tones

Abdominale examination: distended, palpable, not tender or painful, Murphy and Blumberg negative, peristalsis present

Lower limbs: no peripheral edema; peripheral pulses hypophygmic, bilaterally palpable up to the popliteal artery, with no pedidum pulse palpable bilaterally; dry necrosis left first toe, mild cyanosis in the remaining toes of the left foot; no sensory or motor deficits in the left lower limbs  
ECG unchanged from the previous one

Answer: unknown

{8 more few-shot samples - skipped here for readability}

Now evaluate the following:

Category: chronic pulmonary disease  
Options: ['certainly chronic' 'possibly chronic' 'certainly not chronic' 'unknown']

possibly chronic' 'certainly not chronic' 'unknown']

Clinical Text: 77-year-old patient.

Past Medical History: cognitive impairment; lives with RELATIVE and RELATIVE; hearing loss. arterial hypertension, previous Guillain-Barr syndrome in 2007 spinal stabilization surgery in 1993, previous hysterectomy for fibroids, hiatal hernia, cholecystectomy  
SAD, Last geriatric visit  
PHONE\_NUMBER: MRI cerebral vasculopathy - presence of buccal dyskinesias - slow walking with risk of falls  
Therapy: pantoprazole 40-olmesartan 1/2- xanax 0.25- contramal 4 drops x2, levopraid 1/2 x 2, folic acid , stilnox,  
RELATIVE PHONE\_NUMBER

Recent Medical History: patient arrived last night for an episode of sudden illness characterized by a scream followed by generalized TC tremors in all 4 limbs; this morning a second fit was seen in the Emergency and Acceptance Department, TC in all 4 limbs, preceded by a scream.  
Performed in the Emergency Department:  
blood tests: no abnormalities (normal electrolytes), no fever; experiencing sleep loss due to frequent nighttime awakenings. Also performed brain CT brain without contrast which showed a picture of discrete diffuse cortico-subcortical atrophy, no acute manifestations; known chronic vascular disease.

Neurological Examination: patient alert, cooperative, rather quiet, partially oriented in time/place; no clear oculomotor deficits, blinks at bilateral visual threat, no underleveling at Mingazzini I; raises lower limbs alternately for 4-5 sec with symmetrical controlled descent. No appreciable sensory deficits. Bilateral plantar reflex in flexion.

Conclusions: generalized TC seizures (2 episodes) due to moderate cerebral atrophy and cognitive impairment.

Recommendation to start gardenale 50

mg 1 tablet in the evening.  
Clinical observation until tomorrow morning; if there are no other episodes or other illnesses, discharge tomorrow with indication to perform EEG with priority B and check-up in 6 months at epilepsy clinic with blood phenobarbital levels, a complete blood count, and liver and kidney function tests..

Answer:

# Polimi at CRF Filling 2026: Prompt-Based Information Extraction from Italian Clinical Notes

Vittorio Torri<sup>1</sup>, Francesca Ieva<sup>1,2</sup>

<sup>1</sup>MOX - Modelling and Scientific Computing Lab, Department of Mathematics, Politecnico di Milano,

<sup>2</sup>HDS - Health Data Science Centre, Human Technopole

<sup>1</sup>Piazza Leonardo da Vinci, 32, 20133, Milano, Italy,

<sup>2</sup>Viale Rita Levi Montalcini, 1, 20157, Milano, Italy

{vittorio.torri, francesca.ieva}@polimi.it

## Abstract

In this paper we describe the system developed by the Polimi team for the CRF Filling Shared Task 2026, which focuses on extracting structured variables from clinical notes. The task is challenging due to scarce annotations, heterogeneous clinical language, and the sparsity of the 134 items to be extracted. Our approach relies on prompt-based information extraction using locally deployed open-weight Large Language Models (LLMs). We focused on the Italian subset of the dataset. The pipeline performs zero-shot extraction using task-specific prompts augmented with a glossary of abbreviations derived from unlabelled notes. To improve reliability and reduce hallucinations, the extraction schema is decomposed into multiple prompts targeting groups of variables, whose outputs are merged and refined through deterministic post-processing rules to normalize values and recover missing labels. During development we explored verification stages based on LLM-based prediction validation and synthetic example generation, but these strategies did not improve performance and were not included in the final system. On the development set, the best configuration based on Mistral Small 3.2 24B Instruct achieved an F1-score of 67.51%. On the official test set, our system ranked third overall and second among systems evaluated on the Italian subset, achieving an F1-score of 63%.

**Keywords:** Information Extraction, Clinical Natural Language Processing, Large Language Models, Prompt-based Learning, Electronic Health Records

## 1. Introduction

Electronic Health Records (EHRs) contain large amounts of clinically relevant information in the form of unstructured textual documents such as clinical notes and discharge summaries. Automatically extracting structured information from these narratives can support clinical research, hospital management, and decision support systems. However, clinical text processing remains challenging due to the heterogeneity of medical language, the frequent use of abbreviations and implicit expressions, and the presence of context-dependent information (Wang et al., 2018; Locke et al., 2021).

A major limitation in this domain is the scarcity of annotated datasets. Creating high-quality annotations requires medical expertise and strict privacy safeguards, making large annotated corpora difficult to obtain (Wei et al., 2018; Sylolypavan et al., 2023), especially in languages other than English (Névéol et al., 2018). As a result, many clinical NLP tasks must operate in settings with very limited supervision.

Recent advances in Large Language Models (LLMs) have shown promising results for zero- and few-shot clinical information extraction (Hu et al., 2024; Luo et al., 2024; Fornasiere et al., 2024). However, the use of proprietary models in healthcare contexts is often restricted because clinical data cannot be sent to external servers due to pri-

vacancy and regulatory constraints. This motivates the exploration of locally deployable open-weight models that can operate entirely within secure hospital infrastructures.

In this work we describe the system developed by the Polimi team for the CRF Filling Shared Task 2026 (Ferrazzi et al., 2026), which focuses on extracting structured variables from Italian clinical notes. Our approach relies on prompt-based extraction with open-weight LLMs combined with additional knowledge derived from unlabelled data and deterministic post-processing to improve reliability on some specific items.

## 2. Materials and Methods

### 2.1. Data

The dataset provided for the CRF Filling Shared Task consists of short clinical notes, with original Italian texts and corresponding English translation. The aim of the task is to extract 134 items that compose the Case Report Form (CRF), which include symptoms, vital signs, medical history, diagnostic tests, and treatments. These items are divided into binary, categorical, and measured. The main challenge lies in their sparsity:  $\sim 95\%$  of items have *unknown* value. Our team focused on the Italian version of the notes.

The data are divided into a training set (10



labelled notes), a development set (80 labelled notes), a test set (200 notes for which labels have not been released) (Kaczmarek et al., 2026), and a larger collection of 2,667 unlabelled notes (Ferrazzi et al., 2026). In our experiments, the training notes were used exclusively to construct examples for few-shot extraction and for LLM-based verification. The 80 development notes were used during system development to evaluate different prompting strategies and post-processing configurations. The 200 test notes were used only to generate the final predictions submitted to the shared task.

The unlabelled notes were exploited for corpus-level analysis, in particular to identify frequent medical abbreviations and acronyms that were used to construct a glossary integrated into the extraction prompts (Section 2.2.3).

An additional set of 80 synthetic notes (Ferrazzi et al., 2025) provided by the organizers was not used, as their annotated variables were not fully aligned with those in the main dataset.

## 2.2. Methods

Our final submitted system consists of a prompt-based zero-shot extraction pipeline with glossary augmentation and deterministic post-processing for certain items, summarized in Figure 1. In addition, during development we experimented with few-shot learning and additional verification stages designed to filter unsupported predictions and recover potentially missed labels. Although these verification strategies improved precision in some cases, they decreased the overall macro F1-score on the development set and were therefore not included in the final system.

### 2.2.1. Model Selection

We evaluated several open-weight large language models, focusing on models that can be deployed locally without sending clinical data to external servers, a common requirement in hospital environments. In particular, we experimented with LLaMA 3.1 8B, LLaMA 3.3 70B FP8 (Grattafiori et al., 2024), Qwen 2.5 72B AWQ (Yang et al., 2025), and Mistral-Small-3.2-24B-Instruct (Jiang et al., 2023), including also an ensemble of these models. For reference, we additionally included a comparison with two closed-source models from OpenAI, namely GPT-4.1 mini (Achiam et al., 2023) and GPT-5.4 (Singh et al., 2025).

### 2.2.2. Prompt-based Extraction

Information extraction was performed using prompt-based generation. The model receives the clinical note together with detailed instructions specifying

the CRF schema and is asked to return a structured JSON object containing the extracted values. Basic prompts specified only the task and the admissible values, while more complex ones included additional instructions for handling missing values and numerical values.

To reduce hallucinations caused by large output schemas, we decomposed the extraction task into multiple prompts. Each prompt focuses on a subset of clinically related fields (e.g., binary, categorical and measured items). The resulting partial JSON outputs are then merged into a single structured representation before entering the verification stage.

Few-shot prompts included examples derived from the training set to illustrate the expected output format and the exact-match behavior required. Details of the prompts are reported in Appendix B.

### 2.2.3. Glossary Construction from Unlabelled Data

Clinical notes frequently contain abbreviated medical terminology that can hinder reliable information extraction. To address this issue, we constructed a glossary directly from the unlabelled corpus, with the goal of complementing the model’s knowledge in cases where abbreviations are ambiguous or poorly recognized. Acronyms were automatically extracted from the notes using a regular expression identifying tokens composed of two to six upper-case letters. These candidates were aggregated across the corpus and ranked by frequency, resulting in a list of the 1,000 most frequent acronyms.

Each acronym was then processed using an LLM-based interpretation step. The model was prompted to determine whether the acronym corresponds to a well-defined abbreviation in Italian medical language and to provide its meaning together with a confidence score in structured JSON format. We focused on acronyms for which the model exhibited uncertainty or produced incorrect interpretations. Specifically, acronyms associated with low confidence scores ( $< 0.20$ ) or incorrect meanings were incorporated into the extraction prompts as part of a glossary of domain-specific abbreviations. The list of included acronyms with their extracted meaning and confidence is reported in Appendix D.

This design explicitly targets cases where the model’s internal knowledge is unreliable, providing additional context to improve interpretation during extraction. In the dataset, abbreviations are frequent and often correspond to clinical shorthand (e.g., ECG findings, laboratory values, or procedural notes), making their correct interpretation critical for accurate information extraction.

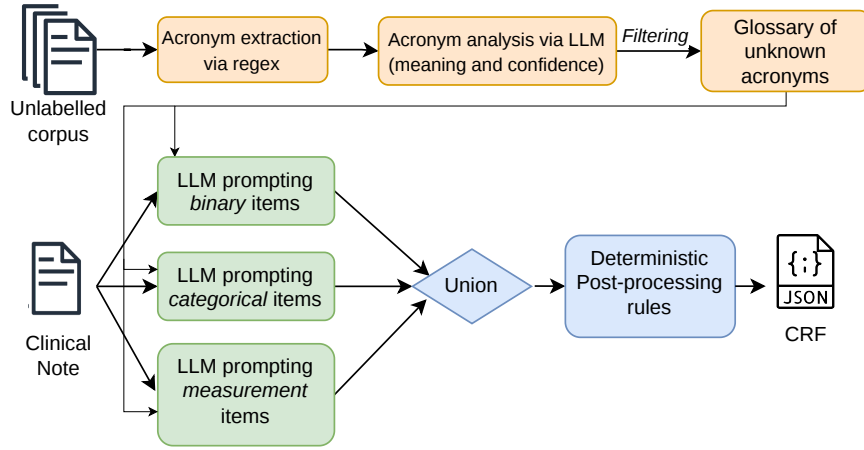


Figure 1: Overview of the extraction pipeline. The clinical note is processed by multiple prompts targeting different subsets of CRF fields. Each of them receives relevant elements from the glossary built from the unlabelled corpus. The resulting partial JSON outputs are merged into a single structured representation and subsequently refined through deterministic post-processing.

#### 2.2.4. Verification Stage

During development we investigated whether an additional verification stage could improve extraction reliability.

First, we evaluated a simple LLM-based verifier that receives the clinical note together with the predicted CRF values and checks whether each prediction is supported by explicit textual evidence. Predictions that cannot be supported by a literal span in the note are discarded and replaced with `unknown`.

In a second variant, we introduced a synthetic-example-guided verifier designed to recover labels initially predicted as `unknown`. To support this step, we generated a bank of synthetic clinical examples using an LLM. Each example consists of a short Italian Emergency Department note in which the value of a specific CRF item is controlled and supported by an explicit evidence span. Generation prompts include stylistic constraints to reproduce realistic ED documentation and information to ensure that the correct label can be determined only from explicit textual evidence. An additional validation pass verifies that the generated note supports the intended label and that the evidence span appears verbatim in the text.

During verification, a small number of contrastive synthetic examples are included in the prompt to guide the model in identifying explicit evidence in the original note. A recovered label is accepted only if a supporting span is explicitly identified.

#### 2.2.5. Post-processing

The predictions generated by the LLM were further refined using a deterministic post-processing

stage applied offline to the model outputs. These rules exploit explicit textual evidence in the clinical notes to correct systematic errors produced by the LLM. The post-processing stage was designed to correct frequent normalization issues and reduce unsupported predictions while preserving the original extraction outputs whenever no clear textual evidence was available. More details, including an example, are in Appendix C.

#### 2.2.6. Ensemble Strategy

We explored an ensemble combining LLaMA, Qwen, and Mistral using a deterministic item-level aggregation. Agreement between at least two models is taken as the final prediction. When only one model predicts a non-`unknown` value, it is accepted only if produced by Mistral, which achieved the highest recall on the development set. In case of conflicting non-`unknown` predictions, Mistral is used as the anchor model. This design leverages agreement as a proxy for confidence while prioritizing recall-oriented predictions.

#### 2.2.7. Implementation Details

All experiments were conducted using locally deployed open-weight models, except those including OpenAI models that were performed by using the official OpenAI APIs. Experiment with Mistral-22B were executed on a single NVIDIA H100 GPU, while smaller models on a single A100 or V100 GPU. LLM inference was performed with the `vLLM` framework (Kwon et al., 2023), which enables batched inference and structured output decoding. Model outputs were constrained to JSON schemas using structured decoding in order to guarantee valid

CRF outputs. Extraction and verification stages were executed with deterministic decoding (temperature = 0) to ensure reproducibility. Code is available at <https://github.com/vittot/CRF-Filling-SharedTask-CL4Health26>.

### 3. Results

Table 1 summarizes the main experiments and the their results on the development set. Baseline zero-shot prompting with Llama 3.1 8B achieved an F1-score of 52.84 %, which increased to 57.78 % when scaling to Llama 3.3 70B FP8. The introduction of rules in the prompt increased significantly the FP, and the glossary significantly reduced them, while keeping a small improvement on TP and FN compared to the first simple prompt. Substantial reduction of FP was obtained when splitting the prompts, with 3 prompts (binary, categorical, measured) being the most effective, even if this also increased significantly the FN. The post-processing steps determined a major improvement, keeping FP at the same level, but reducing FN.

The best development result was obtained with Mistral Small 3.2 24B using three prompts, glossary and enhanced post-processing, reaching an F1-score of 67.51 %. This configuration outperformed both larger but quantized models and ensemble approaches, achieving the best balance between FP and FN. Additional strategies explored during development, including few-shot prompting, ensemble predictions, and verification stages, did not improve the best configuration. In particular, the LLM-based verifier substantially reduced performance, while the synthetic-example-based verifier achieved good F1-score but remained below the best configuration.

Proprietary OpenAI models showed a clear tendency toward over-generation, with a substantially higher number of FP compared to open-weight models (e.g., GPT-5.4: 305 FP vs 94 for Mistral). Despite their good recall, this lower precision resulted in lower F1-score, indicating weaker adherence to the strict evidence-based extraction constraints of the task.

Figure 2 reports the official test set results. The best performance was again obtained by the Mistral-based configuration with three prompts, glossary and enhanced post-processing, achieving an F1-score of 63 %. Systems based on Llama 3.3 70B and Qwen 72B obtained similar but slightly lower results (61 %). The configuration including the verification stage performed worse (53 %), confirming the trends observed on the development set.

Item-level results (Table 2) show substantial variability across variables. Vital signs and laboratory measurements achieved the highest performance,

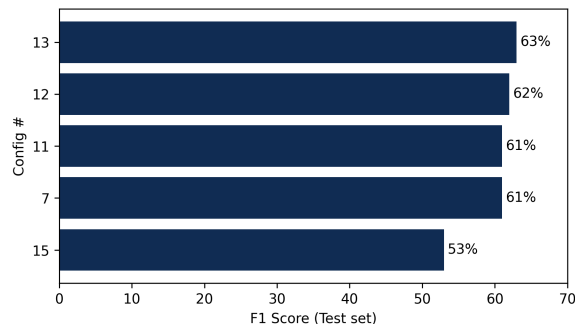


Figure 2: Results on the official test set

often exceeding 85% F1 (e.g., *SpO2*, *heart rate*, *body temperature*, and *blood pressure*), reflecting their relatively standardized textual expressions. Intermediate performance was observed for clinical context variables such as *poly-pharmacological therapy*, *anticoagulants* or *antiplatelet therapy*, and *cardiovascular diseases*. In contrast, symptoms and infrequent conditions showed much lower performance; for example, *presence of dyspnea* achieved an F1-score of 27 %, while several low-support variables obtained F1-scores close to zero.

The best development configuration produced 255 errors, including 155 false negatives and 94 false positives, indicating that the system often failed to extract information explicitly present in the text and predicted `unknown`. False positives were frequently caused by contextual inference (e.g., *cardiovascular diseases*, *poly-pharmacological therapy*), while false negatives were more common for symptoms or episodic events such as *agitation*, *presence of dyspnea*, and *administration of fluids*. Value mismatches mainly involved normalization issues for numerical variables or minor formatting differences (e.g., “hb 12” vs. “12”), suggesting that improved normalization rules could further reduce these errors. Some examples are reported in Appendix A.

### 4. Discussion

In this work we explored several strategies for prompt-based information extraction from clinical notes, including zero-shot and few-shot prompting, glossary-based knowledge injection, LLM-based verification, and deterministic post-processing. While scaling to larger models improved baseline performance, prompt engineering alone produced modest gains. The non-quantized Mistral Small 3.2 24B model outperformed both the smaller Llama 3.1 8B and the larger but quantized Llama 3.3 70B, suggesting that architecture and quantization may have a stronger impact than parameter count alone. Additional strategies yielded limited improvements (glossary) or no gains (verification). In contrast,

| Config # | Model                 | Description                                                                             | TP         | FP        | FN        | F1 [%]       |
|----------|-----------------------|-----------------------------------------------------------------------------------------|------------|-----------|-----------|--------------|
| 1        | Llama 3.1 8B          | simple prompt                                                                           | 235        | 337       | 107       | 52.84        |
| 2        | Llama 3.3 70B FP8     | simple prompt                                                                           | 306        | 380       | 60        | 57.78        |
| 3        | Llama 3.3 70B FP8     | prompt with rules                                                                       | 309        | 760       | <b>50</b> | 56.21        |
| 4        | Llama 3.3 70B FP8     | rules + glossary                                                                        | 310        | 498       | 54        | 58.23        |
| 5        | Llama 3.3 70B FP8     | 3 prompts rules + glossary                                                              | 148        | 88        | 234       | 57.76        |
| 6        | Llama 3.3 70B FP8     | 3 prompts rules + glossary + post-proc                                                  | 231        | 90        | 136       | 64.13        |
| 7        | Llama 3.3 70B FP8     | 3 prompts rules + glossary + enh. post-proc                                             | 245        | 100       | 121       | 65.30        |
| 8        | Llama 3.3 70B FP8     | 3 prompts few-shot rules + glossary + post-proc                                         | 210        | <b>87</b> | 157       | 62.31        |
| 9        | Qwen 2.5 72B AWQ      | 3 prompts rules + glossary                                                              | 146        | 92        | 227       | 58.48        |
| 10       | Qwen 2.5 72B AWQ      | 3 prompts rules + glossary + post-proc                                                  | 229        | 108       | 129       | 63.16        |
| 11       | Qwen 2.5 72B AWQ      | 3 prompts rules + glossary + enh. post-proc                                             | 248        | 103       | 114       | 65.66        |
| 12       | Mistral Small 3.2 24B | 3 prompts rules + glossary + post-proc                                                  | 235        | 96        | 130       | 64.36        |
| 13       | Mistral Small 3.2 24B | 3 prompts rules + glossary + enh. post-proc                                             | 255        | 94        | 155       | <b>67.51</b> |
| 14       | Ensemble              | Llama70B-FP8 + Qwen72B-AWQ + MistralSmall24B                                            | 238        | 121       | 138       | 63.20        |
| 15       | Mistral Small 3.2 24B | 3 prompts + rules + glossary + post-proc + LLM-based verifier (Llama 3.1 8B)            | 302        | 498       | 64        | 58.23        |
| 16       | Mistral Small 3.2 24B | 3 prompts + rules + glossary + post-proc + Synth-based verifier (Mistral Small 3.2 24B) | 266        | 133       | 101       | 65.97        |
| 17       | OpenAI GPT 4.1 mini   | 3 prompts + rules + glossary                                                            | 288        | 885       | 95        | 55.06        |
| 18       | OpenAI GPT 5.4        | 3 prompts + rules + glossary                                                            | 307        | 317       | 82        | 62.43        |
| 19       | OpenAI GPT 4.1 mini   | 3 prompts + rules + glossary + enh. post-proc                                           | 301        | 867       | 81        | 57.55        |
| 20       | OpenAI GPT 5.4        | 3 prompts + rules + glossary + enh. post-proc                                           | <b>314</b> | 305       | 75        | 64.22        |

Table 1: Development set results for the different configurations of our system.

| Item                                  | F1   | Support |
|---------------------------------------|------|---------|
| SpO2                                  | 1.00 | 20      |
| Heart rate                            | 0.91 | 25      |
| Body temperature                      | 0.90 | 28      |
| Blood pressure                        | 0.86 | 24      |
| Level of consciousness                | 0.88 | 33      |
| Poly-pharmacological therapy          | 0.79 | 23      |
| Anticoagulants / antiplatelet therapy | 0.81 | 16      |
| Antihypertensive therapy              | 0.81 | 19      |
| Cardiovascular diseases               | 0.74 | 27      |
| Chronic pulmonary disease             | 0.73 | 7       |
| Neuropsychiatric disorders            | 0.46 | 10      |
| ECG abnormalities                     | 0.44 | 7       |
| Level of autonomy (mobility)          | 0.33 | 11      |
| Presence of dyspnea                   | 0.27 | 20      |

Table 2: Item-level performance of the best development configuration on selected items.

decomposing the task into three prompts targeting groups of related variables proved beneficial, likely reducing hallucinations caused by the large output schema. This observation is consistent with recent studies reporting limited benefits of more complex techniques such as RAG or CoT in clinical NLP (Nagar et al., 2025). Deterministic post-processing further improved performance by correcting normalization errors and recovering information expressed through predictable textual patterns. Using this configuration, our system ranked third overall in the shared task and second among systems evaluated on the Italian subset.

Performance varied substantially across variables. Vital signs and laboratory measurements were extracted reliably due to their stable lexical patterns, whereas symptoms, contextual variables, and rarely mentioned conditions remained challeng-

ing because of implicit phrasing, abbreviations, and limited supervision.

This work has several limitations. The dataset is relatively small, limiting the ability to learn robust patterns for rare variables and increasing the risk of overfitting the development set. Experiments were conducted only on Italian notes, and the generalizability to the English version remains to be investigated. Moreover, the best-performing models require computational resources that may not always be available in hospital environments.

Future work could explore hybrid approaches combining prompt-based extraction with lightweight supervised models, improved normalization and abbreviation handling, the use of English prompts on Italian data, and larger multilingual clinical corpora. A more systematic evaluation of synthetic data generation for supervision may also help improve performance on sparse variables.

## 5. Acknowledgements

The present research has been partially supported by the Italian Ministry of University and Research (MUR), grant Dipartimento di Eccellenza 2023-2027, awarded to the Department of Mathematics, Politecnico di Milano.

## 6. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florenzia Leoni Aleman, Diogo Almeida, Janko Al-



- tenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Pietro Ferrazzi, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. 2026. Overview of the crf 2026 shared task on clinical case report forms filling. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma, Mallorca (Spain). ELRA.
- Raffaello Fornasiere, Nicolò Brunello, Vincenzo Scotti, and Mark Carman. 2024. Medical information extraction with large language models. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 456–466.
- 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*.
- Danqing Hu, Bing Liu, Xiaofeng Zhu, Xudong Lu, and Nan Wu. 2024. Zero-shot information extraction from radiological reports using chatgpt. *International Journal of Medical Informatics*, 183:105321.
- Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b. arxiv. *arXiv preprint arXiv:2310.06825*, 10:3.
- 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.
- Saskia Locke, Anthony Bashall, Sarah Al-Adely, John Moore, Anthony Wilson, and Gareth B Kitchen. 2021. Natural language processing in medicine: a review. *Trends in Anaesthesia and Critical Care*, 38:4–9.
- Xiao Luo, Fattah Muhammad Tahabi, Tressica Marc, Laura Ann Haunert, and Susan Storey. 2024. Zero-shot learning to extract assessment criteria and medical services from the preventive healthcare guidelines using large language models. *Journal of the American Medical Informatics Association*, 31(8):1743–1753.
- Aishik Nagar, Viktor Schlegel, Thanh-Tung Nguyen, Hao Li, Yuping Wu, Kuluhan Binici, and Stefan Winkler. 2025. Llms are not zero-shot reasoners for biomedical information extraction. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 106–120.
- Aur lie N v  ol, Hercules Dalianis, Sumithra Velupilai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.
- Aneeta Syloypavan, Derek Sleeman, Honghan Wu, and Malcolm Sim. 2023. The impact of inconsistent human annotations on ai driven clinical decision making. *NPJ Digital Medicine*, 6(1):26.
- Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.
- Qiang Wei, Amy Franklin, Trevor Cohen, and Hua Xu. 2018. Clinical text annotation—what factors are associated with the cost of time? In *AMIA Annual Symposium Proceedings*, volume 2018, page 1552.
- 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*.

## 7. Language Resource References

- Pietro Ferrazzi, Mattia Franzin, Alberto Lavelli, and Bernardo Magnini. 2026. [Small llms for medical nlp: a systematic analysis of few-shot, constraint decoding, fine-tuning and continual pre-training in italian](#).
- Pietro Ferrazzi, Alberto Lavelli, and Bernardo Magnini. 2025. [Converting annotated clinical cases into structured case report forms](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 307–318, Viena, Austria. Association for Computational Linguistics.



Gabriela Anna Kaczmarek, Pietro Ferrazzi,  
Lorenzo Porta, Vicky Rubini, and Bernardo  
Magnini. 2026. [Toward automatic filling of case  
report forms: A case study on data from an  
italian emergency department.](#)

## A. Evaluation Details and Examples

The task is evaluated at the item level across clinical documents, using macro-averaged F1-score as the official metric. For each document, systems must predict the values of a predefined set of CRF items, which are compared with the reference annotations. The value *unknown*, allowed for all items, indicates that the information is not explicitly present in the document. An item contributes to the evaluation whenever at least one of the two values (prediction or reference) is different from unknown. Correct matches are counted as true positives (TP), while predicting a value where the reference is *unknown* produces a false positive (FP) and predicting *unknown* where a value exists produces a false negative (FN).

Macro-F1 is computed across item values for each document and then averaged across all documents to obtain the final score. This design gives equal importance to each document and prevents documents with many populated items from dominating the metric. At the same time, rare items and frequent items contribute equally, making robust performance across all CRF fields important for achieving a high overall score.

The official evaluation script computes the macro-F1, TP, FP and TN, and it was used to report results on development set in Table 1, while the evaluation on the test is on the submission platform which reports only the macro-F1 (Figure 2).

In addition to the official scoring script, we performed a per-item error analysis on the development set, to better understand model behavior (summarized in Table 2). We also manually investigated mistakes on specific document/item pairs, and we report here some examples, including also potential annotation mistakes.

### Example 1 (Chest Pain - FN - Terms not recognized)

**Document ID:** 419929

**Item:** chest pain

**Gold value:** y

**Prediction:** unknown

**Clinical note excerpt:**

*"Riferisce **algie** ben localizzate all'**emitorace** di sinistra da una settimana."*

The model is not recognized the presence of chest pain, perhaps due to the word *emitorace* (*hemithorax*) which indicates only half of the chest.

### Example 2 (Respiratory Rate - FP - Suspect Annotation Mistake)

**Document ID:** 1498690

**Item:** respiratory rate

**Gold value:** unknown

**Prediction:** eupneic

**Clinical note excerpt:**

*"B: **eupnoico** in cn 2 l/min, lieve utilizzo muscolatura accessoria, e/to ega art.→SO2 93%"*

The text clearly states that the patient is *eupneic*, so the gold label is probably wrong.

### Example 3 (SPO2 - FP - Normalization error)

**Document ID:** 1777967

**Item:** SPO2

**Gold value:** 97%

**Prediction:** 97%aa

**Clinical note excerpt:**

*"PAOS 120/75, FC 115, **Sat 97%aa**, T 38.1°C"*

This error is likely due to the instruction given to the LLM not to normalize measurement values. In the annotations, measurement fields often preserve spaces and abbreviations as they are in the original text (e.g., *hb 12, 97%*), which are kept as part of the gold value, apparently not in a fully consistent manner.

### Example 4 (History of Allergy - FP - Negation not recognized)

**Document ID:** 1307609

**Item:** History of Allergy

**Gold value:** n

**Prediction:** y

**Clinical note excerpt:**

*"**Non fumo né allergie note.**"*

The model recognized the mention of allergy (it is not *unknown*), but it missed the negation, expressed in a non-trivial form which also bypasses our dedicated post-processing rule.

### Example 5 (ECG Abnormality - FN - Acronym not recognized)

**Document ID:** 821011

**Item:** ECG Abnormality

**Gold value:** y

**Prediction:** unknown

**Clinical note excerpt:**

*"- ECG → RS FC 58 asse a sx, 1 **CPV**"*

The model does not seem to recognize the mention of CPV, which stands for *Premature Ventricular Contraction* and indicates an abnormality in the ECG.

### Example 6 (Neurodegenerative Diseases - FN - Borderline condition)

**Document ID:** 382962

**Item:** Neurodegenerative Diseases

**Gold value:** y

**Prediction:** unknown

**Clinical note excerpt:**

“- **Sindrome di Arnold Chiari**. Seguita c/o H Molinette. In programma RM encefalo a breve (non ancora effettuata) e successiva v NCH con dr. NOME\_PERSONA per presa in carico presso OSGB.”

The model does not recognize the mention of *Sindrome di Arnold Chiari*, a neurological disorder reported in the patient's history. Although not strictly a neurodegenerative disease (e.g., Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis), this mention may still be relevant depending on how neurological conditions are interpreted within the annotation schema.

## B. Prompt Templates

This appendix reports the prompt templates used in our extraction pipeline. Prompts were designed to extract structured variables from Italian clinical notes using a zero-shot instruction-following setup. We also include prompts for verifiers and glossary analysis.

### B.1. Binary Variables Prompt

Sei un estraattore di dati clinici per verbali di Pronto Soccorso italiani.

OBIETTIVO: Compila SOLO i campi binari (y/n/unknown) richiesti.

REGOLE CRITICHE (ANTI-FALSE-POSITIVE):

- 1) DEFAULT = "unknown". Se il concetto NON e' esplicitamente menzionato, scrivi "unknown".
- 2) Scrivi "n" SOLO con negazione esplicita nel testo ("nega", "non", "assente", "negativo", "non segni di", ecc.).
- 3) Scrivi "y" SOLO con affermazione esplicita nel testo.
- 4) NON fare inferenze cliniche, NON dedurre diagnosi o terapie non dichiarate.
- 5) Filtra temporalmente: considera SOLO dati all'ADMISSION/triage.

ECCEZIONE TEMPORALE:

Per item di ANAMNESI/TERAPIA DOMICILIARE/SOCIALE e' consentito usare sezioni

"APR:", "Anamnesi:", "TD:", "Terapia domiciliare:", "Allergie:", o elenchi

di farmaci/patologie come evidenza esplicita.

- 6) **\*\*Administration of [Drugs]\*\***: Scrivi 'y' solo se il farmaco e' stato

somministrato FISICAMENTE in PS (verbi: "praticato", "somministrato", "fatto").

I farmaci presenti in TD NON contano come 'administration of ...'.

ECCEZIONE SOLO PER ITEM DI

ANAMNESI/TD/SOCIALE (riduzione FN controllata):

Per i seguenti item:

- history of allergy
- poly-pharmacological therapy
- antihypertensive therapy
- cardiovascular diseases
- anticoagulants or antiplatelet drug therapy
- diffuse vascular disease
- neuropsychiatric disorders
- living alone

vale questa regola speciale:

- A) Se trovi una sezione tipo "APR:", "Anamnesi:", "TD:", "Terapia domiciliare:", "Allergie:" oppure un elenco di farmaci/patologie separati da virgole o trattini, questo conta come MENZIONE ESPLICITA quindi puoi scrivere "y" se pertinente.
- B) Scrivi "n" SOLO con negazione esplicita (es. "nega allergie", "allergie: neg", "non assume terapia", ecc.).
- C) Se non trovi queste sezioni o elenchi espliciti allora resta "unknown".
- D) NON applicare questa eccezione agli altri item binari (acuti).

GLOSSARIO (serve solo per riconoscere menzioni esplicite, non per dedurre):

- **\*\*EON\*\***: Esame Obiettivo Neurologico. Se scritto "EON nella norma" o "GCS 15" o "vigile/orientato" allora level of consciousness = 'a'. Se non scritto allora "unknown".
- **\*\*CE\*\***: Corpo estraneo (utile per 'foreign body in the airways').
- **\*\*FA\*\***: Fibrillazione atriale (utile per 'arrhythmia').
- **\*\*NAO / TAO\*\***: terapia anticoagulante (utile per 'anticoagulants or antiplatelet drug therapy').
- **\*\*ASA / aspirina / clopidogrel\*\***: antiaggreganti (utile per 'anticoagulants or antiplatelet drug therapy').
- **\*\*TD\*\***: Terapia domiciliare.

REGOLE PER OGNI ITEM:

{rules\_text}

Restituisci ESCLUSIVAMENTE un oggetto JSON conforme allo schema.

## B.2. Categorical Variables Prompt

Sei un estrattore di dati clinici per verbali di Pronto Soccorso italiani.

OBIETTIVO: Compila SOLO i campi categorici richiesti (non misurazioni).

REGOLE:

- 1) DEFAULT = "unknown" se l'informazione non e' esplicitamente nel testo.
- 2) Non dedurre "n" o altre categorie se non scritto.
- 3) Per ogni campo, usa ESATTAMENTE una delle opzioni permesse.
- 4) Considera SOLO l'ADMISSION/triage (ignora rivalutazioni).
- 5) **\*\*Level of Consciousness\*\***: 'a' (vigile), 'v' (verbal), 'p' (pain), 'u' (unresponsive). 'a' SOLO se il testo contiene esplicitamente: "vigile", "orientato", "GCS 15", "EON nella norma". Altrimenti "unknown".

GLOSSARIO (Il glossario serve solo a riconoscere menzioni esplicite. Non usarlo per dedurre diagnosi non scritte):

- **\*\*APR\*\***: Anamnesi Patologica Remota (storia clinica passata).
- **\*\*EON\*\***: Esame Obiettivo Neurologico (stato di coscienza). Se "EON nella norma" -> level of consciousness = 'a'.
- **\*\*AASS\*\*** / **\*\*AAII\*\***: Arti Superiori / Arti Inferiori.
- **\*\*TC\*\***: Se associata a gradi (es. TC 37.1) e' **\*\*Temperatura Corporea\*\***. Se in contesto radiologico e' Tomografia Computerizzata.

REGOLE PER OGNI ITEM:  
{rules\_text}

Restituisci ESCLUSIVAMENTE un oggetto JSON conforme allo schema.

## B.3. Measured Variables Prompt

Sei un estrattore di dati clinici per verbali di Pronto Soccorso italiani.

OBIETTIVO: Compila SOLO i campi MISURATI richiesti (valori numerici/testuali).

REGOLE MIRROR (EXACT MATCH):

- 1) Se un valore e' presente nel testo, COPIALO LETTERALMENTE (stessi spazi, simboli, unita' se presenti, ad esempio "hb 12", "k 6.3").
- 2) Non normalizzare (es: non trasformare "99 %" in "99%").
- 3) Se un valore non e' presente, scrivi "unknown".
- 4) Considera SOLO valori all'ADMISSION/triage.
- 5) Per 'spo2' scrivi SOLO la percentuale cosi' come appare (es. "99 %" o "99%"), senza prefissi ("SpO2", "Sat", ecc.).

GLOSSARIO (Il glossario serve solo a riconoscere menzioni esplicite. Non usarlo per dedurre diagnosi non scritte):

- **\*\*APR\*\***: Anamnesi Patologica Remota (storia clinica passata).
- **\*\*HGT\*\***: Glicemia capillare (valore del glucosio).
- **\*\*EGA\*\***: Emogasanalisi (fonte per pH, pO2, pCO2, Lattati).
- **\*\*AA\*\***: Aria Ambiente (SpO2 senza ossigeno).
- **\*\*ASA\*\***: Acido Acetilsalicilico.
- **\*\*NRS\*\*** / **\*\*VAS\*\***: Scale del dolore (0-10) ..
- **\*\*TC\*\***: Se associata a gradi (es. TC 37.1) e' **\*\*Temperatura Corporea\*\***. Se in contesto radiologico e' Tomografia Computerizzata.
- **\*\*GB / WBC\*\***: Sono i **\*\*Leucociti\*\*** (Globuli Bianchi). Se leggi "gbb 16", scrivi "gbb 16" nel campo leukocytes.

REGOLE PER OGNI ITEM:  
{rules\_text}

Restituisci ESCLUSIVAMENTE un oggetto JSON conforme allo schema.

## B.4. Acronym Analysis Prompt

Sei un medico esperto di Pronto Soccorso. Analizza l'acronimo medico e rispondi solo in formato JSON seguendo lo schema richiesto.

Analizza l'acronimo: {acronym}.  
E' un termine certo nel contesto dell'emergenza urgenza italiana?

Rispondi con questo schema JSON:

```
{{"acronym": "{acronym}",
 "known": "yes/no",
 "meaning": "...",
 "certainty_score": 0-100}}
```

## B.5. LLM-based Verifier Prompt

Sei un verificatore rigoroso.

Riceverai:

- 1) Una nota clinica di Pronto Soccorso (in italiano).
- 2) Un elenco di variabili del CRF con i valori proposti (predizioni).

Compito:

Per OGNI variabile, devi decidere se il valore proposto e' SUPPORTATO da uno span ESPLICITO nel testo della nota.

Regole:

- "SUPPORTATO" solo se la nota lo afferma esplicitamente (oppure lo nega esplicitamente se il valore e' una negazione come 'n').
- Se la variabile non e' menzionata esplicitamente, segnala "DROP".
- NON utilizzare inferenze cliniche.
- NON assumere valori negativi in assenza di evidenza. Silenzio implica DROP.
- L'output deve essere in formato JSON con chiavi = nomi delle variabili e valori in {"SUPPORTATO", "DROP"}.

## B.6. Synth-based Verifier Prompt

Sei un verificatore clinico per estrazione strutturata da verbali di Pronto Soccorso italiani.

OBIETTIVO:

Per UN SOLO item target, verifica se il valore iniziale "unknown" deve essere mantenuto oppure corretto in base a evidenza ESPLICITA presente nel testo.

REGOLE CRITICHE:

- 1) DEFAULT = "unknown".
- 2) Puoi cambiare "unknown" in "y" o "n" SOLO se trovi evidenza ESPLICITA nel testo.
- 3) NON fare inferenze cliniche.
- 4) NON usare conoscenza medica implicita per dedurre il valore.
- 5) Usa SOLO il testo del verbale.

- 6) L'evidence\_span deve essere una sottostringa LETTERALE del verbale, non parafrasata.
- 7) Se non trovi una sottostringa letterale di supporto, mantieni "unknown".
- 8) Se ci sono elementi contraddittori o poco chiari, preferisci "unknown".

ITEM TARGET:

{item}

DESCRIZIONE ITEM:

{item\_description}

ISTRUZIONI SPECIFICHE:

{item\_rules}

ESEMPI CONTRASTIVI DI RIFERIMENTO:

{examples\_block}

VERBALE CLINICO:

{note}

VALORE INIZIALE DEL PRIMO PASSAGGIO:

unknown

RESTITUISCI ESCLUSIVAMENTE un JSON conforme allo schema.

## B.7. Synthetic Examples Generation Prompt

Sei un generatore di brevi verbali clinici sintetici italiani di Pronto Soccorso.

OBIETTIVO:

Genera un singolo verbale clinico realistico, breve ma plausibile, in italiano, tale che il valore corretto per il campo target sia ESATTAMENTE quello richiesto.

VINCOLI:

- 1) Il verbale deve sembrare realistico e non artificiale.
- 2) Non nominare mai il nome tecnico del campo target in inglese.
- 3) Non inserire spiegazioni, solo il contenuto richiesto.
- 4) Il testo deve contenere abbastanza contesto clinico da sembrare un vero verbale.
- 5) Il valore corretto del campo target deve essere determinabile SOLO da un'evidenza esplicita nel testo.
- 6) Se label = "unknown", NON inserire alcuna evidenza esplicita ne' positiva ne' negativa per il campo target.
- 7) Mantieni lo stile compatibile con



verbali PS italiani: triage, anamnesi, APR, TD, EO, terapia, decorso, ecc.

- 8) Evita di copiare identicamente gli esempi eventualmente forniti.
- 9) Inserisci uno o piu' distrattori realistici ma NON contraddire il label target.
- 10) L'evidenza deve essere locale e verificabile come span testuale.

TARGET ITEM:  
{item}

DESCRIZIONE:  
{description}

LABEL RICHIESTA:  
{label}

INDICAZIONI SU EVIDENZA POSITIVA:  
{positive\_evidence\_hint}

INDICAZIONI SU EVIDENZA NEGATIVA:  
{negative\_evidence\_hint}

INDICAZIONI SU UNKNOWN:  
{unknown\_evidence\_hint}

SEZIONI PROBABILI:  
{likely\_sections}

INDIZI LESSICALI POSSIBILI:  
{lexical\_hints}

{anchors\_block}

ESEMPI:  
  
{few\_shots\_block}

RESTITUISCI ESCLUSIVAMENTE un JSON conforme allo schema.

## C. Post-Processing Rules

After LLM-based extraction, predictions were refined using a deterministic post-processing pipeline applied to the generated JSON outputs.

For each document, the predicted items are first mapped to an item–value dictionary. The clinical note is then analyzed using regular expressions and rule-based heuristics. Rules are applied sequentially and may modify predicted values according to the conditions described below.

Unless otherwise stated, rules modify a prediction only if the current value is `unknown`.

The simple "post-processing" named in the paper cover rules for *history based inference*, *allergy negation* and additional rules for adjusting measurement units of measured variables, while the

others reported here are part of the "enhanced post-processing".

Table C1 reports an example of the post-processing effect on the following document from the dev set (id 914210):

Pte proveniente con ambulanza, preso in carico in sala visita covid per dispnea

A: vie aeree pervie

B: giunge in MR, E/EGA ART in AA, spo2 86% pO2 49 pCO2 29, posizionate cn 4 lt, eupnoico al momento

C: cute asciutta, polsi isosfigmici bilateralmente FC 89 b/min R PAO 130/80

D: GCS 15 pupille isocicliche bilateralmente, vigile orientato

E: tc 38 E/perfalgan in ambulanza

cvp posizionato dal LUOGO in sede E/ecg

LUOGO in degenza covid per nota positivita' al covid

4AT negativa

In this note, post-processing recovers two missed labels (*foreign body in the airways*=*n*, *presence of dyspnea*=*y*) and suppresses two unsupported predictions (*presence of respiratory distress*, *sars-cov-2 swab test*), while preserving all already correctly extracted values. The model originally predicted the *sars-cov-2 swab test* since it is mentioned that the patient comes from *sala visite covid (room covid visits)*. Regarding dyspnea, the text clearly mentions it, so it is surprising it was originally missed. On the contrary, the *presence of respiratory distress=current* was a reasonable inference given the values of spo2, paO2, pacO2, and the oxygen therapy (*posizionate cn 4 lt*), (and the presence of dyspnea), but it is not explicitly mentioned in the text so annotations do not include it. The *foreign body in the airways* can be recognized to be negative due to the mention of *vie aeree pervie (patent airway)*. This is a very common mention in these notes, but the model does not recognize its meaning.

## Dyspnea Recovery Rule

### Positive lexical triggers:

- dispnea
- dispnoe
- fiato corto
- affanno
- fame d'aria
- difficoltà respiratoria

| Item                             | Gold         | Raw          | Post-proc      | Effect             |
|----------------------------------|--------------|--------------|----------------|--------------------|
| level of consciousness           | A            | A            | A              | unchanged, correct |
| respiratory rate                 | eupneic      | eupneic      | eupneic        | unchanged, correct |
| body temperature                 | hyperthermic | hyperthermic | hyperthermic   | unchanged, correct |
| heart rate                       | normocardic  | normocardic  | normocardic    | unchanged, correct |
| blood pressure                   | normotensive | normotensive | normotensive   | unchanged, correct |
| spo2                             | 86%          | 86%          | 86%            | unchanged, correct |
| foreign body in the airways      | n            | unknown      | <b>n</b>       | recovered FN       |
| presence of dyspnea              | y            | unknown      | <b>y</b>       | recovered FN       |
| paO2                             | 49           | 49           | 49             | unchanged, correct |
| pacO2                            | 29           | 29           | 29             | unchanged, correct |
| presence of respiratory distress | unknown      | current      | <b>unknown</b> | removed FP         |
| sars-cov-2 swab test             | unknown      | pos          | <b>unknown</b> | removed FP         |

Table C1: Example of information extraction from a note (914210), including gold labels (Gold) and output from the model before (Raw) and after post-processing (Post-proc).

### Negation patterns:

- `neg dispnea`
- `dispnea negata`
- `non presenta dispnea`
- `assenza di dispnea`
- `senza dispnea`
- `no dispnea`

### Rule:

- If a negation pattern occurs in the note, the value is set to `n`.
- Otherwise, if a dyspnea lexical trigger occurs without a nearby negation (within  $\pm 80$  characters), the value is set to `y`.

### Foreign Body in Airways Rule

#### Positive triggers:

- `corpo estraneo`
- `inalazione di corpo estraneo`
- `aspirazione di corpo estraneo`
- `soffocamento`
- `choking`

#### Negative triggers:

- `vie aeree pervie`

### Rule:

- If a negative trigger occurs, the value is set to `n`.
- Otherwise, if a positive trigger occurs, the value is set to `y`.

### Blood Pressure Normalization

The clinical note is searched for patterns of the form

```
(PA|pressione arteriosa)
<systolic>/<diastolic>
```

where systolic and diastolic are numerical values. The extracted values are categorized as follows:

- **hypertensive:** systolic  $\geq 140$  or diastolic  $\geq 90$
- **hypotensive:** systolic  $< 90$  or diastolic  $< 60$
- **normotensive:** otherwise

If such a pattern is detected, the prediction for the item is replaced with the corresponding category.

### Allergy Negation Rule

The following expressions are interpreted as explicit negation:

- `allergie negate`
- `nessuna allergia`
- `no allergie`
- `nega allergie`
- `NKA`
- `NAD`
- `NKDA`

### Rule:

If one of these expressions is detected, the value is set to `n`. This rule may override both `unknown` and incorrectly predicted positive values.

### Evidence-based Filtering Rules

Some variables frequently produced false positive predictions. For these variables, predictions different from `unknown` are kept only if explicit textual evidence is detected. Otherwise, the value is reset to `unknown`.

### Respiratory Distress

Evidence patterns include:

- `distress respiratorio`
- `insufficienza respiratoria`
- `IRA`
- `tirage`
- `uso dei muscoli accessori`
- `cianosi`

If none of these patterns occurs in the note, the prediction is set to `unknown`.

### Level of Consciousness

Evidence patterns include:

- `vigile`
- `orientato`
- `GCS`
- `cosciente`

If none of these patterns occurs, the prediction is set to `unknown`.

### Level of Autonomy (Mobility)

Evidence patterns include:

- `deambula`
- `cammina`
- `carrozzina`
- `bastone`
- `girello`

If none of these patterns occurs, the prediction is set to `unknown`.

### Chronic Cardiac Failure

Evidence patterns include:

- `scompenso cardiaco`
- `insufficienza cardiaca`

If none of these expressions occurs, the prediction is set to `unknown`.

### SARS-CoV-2 Swab Test

A prediction is retained only if the note contains the word `tampone` together with one of the following:

- `positivo`
- `negativo`
- `eseguito`
- `effettuato`
- `risultato`

Otherwise the value is set to `unknown`.

### Chronic Dialysis

Evidence patterns include:

- `emodialisi`
- `dialisi`
- `dialisi peritoneale`

If none of these expressions occurs, the prediction is set to `unknown`.

### History-based Inference Rules

Additional rules recover missing values from explicit mentions of therapies or comorbidities.

### Anticoagulant or Antiplatelet Therapy

Mentions of the following medications trigger a positive label:

- `warfarin`
- `coumadin`
- `clopidogrel`
- `aspirin`
- `cardioaspirin`
- `apixaban`
- `rivaroxaban`
- `dabigatran`

### Antihypertensive Therapy

Mentions of antihypertensive medications such as ACE inhibitors, ARBs, beta blockers, calcium channel blockers, or diuretics trigger a positive label.

### Cardiovascular Diseases

Evidence patterns include:

- `cardiopatìa`
- `fibrillazione atriale`
- `coronaropatìa`
- `ipertensione`

### Diffuse Vascular Disease

Evidence patterns include:

- vasculopatia
- arteriopatia
- aterosclerosi

### Neuropsychiatric Disorders

Evidence patterns include:

- demenza
- alzheimer
- disturbo psichiatrico
- depressione

### Living Alone

Expressions such as *vive da solo* set the value to *y*. Expressions such as *vive con la moglie* or *vive con la famiglia* set the value to *n*.

### Poly-pharmacological Therapy

If the therapy section of the note contains at least three medication separators (e.g., commas or semi-colons) or at least four drug tokens, the value is set to *y*.

### Diagnostic Exam Rules

Rule-based extraction was also applied to imaging and diagnostic exam variables:

- brain CT scan, any abnormality
- abdomen CT scan, any abnormality
- cardiac ultrasound, any abnormality
- thoracic ultrasound, any abnormalities
- chest RX, any abnormalities
- ECG, any abnormality

For each exam, the algorithm detects whether the procedure is mentioned and whether abnormal findings are reported using terms such as *alterazioni*, *patologico*, or *anomalie*. If the exam is mentioned but no abnormality is detected, the prediction is set to *unknown*.

## D. Glossary

Table D1 reports the acronyms included in the glossary, with the meaning and confidence extracted by the LLM and the true meaning. Except for two cases, all the others are wrong meanings despite high confidence of the LLM.

| Acronym | Known | Conf. | Known Meaning                                                 | True Meaning                                                                                                                   |
|---------|-------|-------|---------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| EON     | no    | 0     | –                                                             | Esame Obiettivo Neurologico (Neurological physical examination)                                                                |
| CE      | yes   | 80    | Codice di Emergenza (Emergency Code)                          | Corpo Estraneo (Foreign Body)                                                                                                  |
| FA      | yes   | 80    | Fibra Ariana, Frattura dell'Anca (Ariana Fibre, Hip Fracture) | Fibrillazione Atriale (Atrial Fibrillation)                                                                                    |
| NAO     | yes   | 80    | Non Applicabile Operativamente (Not applicable in practice)   | Nuovi Anticoagulanti Orali (New Oral Anticoagulants)                                                                           |
| TAO     | yes   | 90    | Trombosi Arteriosa Ostiale (Ostial Arterial Thrombosis)       | Terapia Anticoagulante Orale (Oral anticoagulant therapy)                                                                      |
| TD      | yes   | 80    | Tachicardia Da (Tachicardia Da)                               | Terapia Domiciliare (Home Care)                                                                                                |
| APR     | yes   | 80    | Acid Peptic Reflux                                            | Anamnesi Patologica Remota (Past medical history)                                                                              |
| AASS    | yes   | 100   | Automated External Defibrillator                              | Arti Superiori (Upper Limbs)                                                                                                   |
| TC      | yes   | 90    | Trauma Cranico (Head Injury)                                  | If associated with degrees: Temperatura Corporea (Body Temperature); in radiology context: Tomografia Computerizzata (CT scan) |
| HGT     | no    | 0     | –                                                             | Hemo Gluco Test                                                                                                                |
| EGA     | yes   | 100   | Età Gestazionale Anamnestica (Gestational Age)                | Emogasanalisi (Blood gas analysis)                                                                                             |
| NRS     | yes   | 80    | Numero Registro Sanitario (Number of Health Registry)         | Numerical Pain Rating Scale                                                                                                    |
| GB      | yes   | 80    | Gravità del Bene (Severity of Goodness)                       | Globuli Bianchi (White Blood Cells)                                                                                            |

Table D1: Acronyms of the glossary extracted from the unlabelled corpus, with the meaning and confidence reported by the LLM and the true meaning.



# Cohere Labs Community at FoodBench-QA 2026: The Cake Makes the Ingredients

Ravi Ranjan<sup>1</sup>, Roshan Santhosh<sup>2</sup>, Lucien Carroll<sup>3</sup>

<sup>1</sup>GLA University, India <sup>2</sup>Adobe, USA <sup>3</sup>Kedara, Inc.

Cohere Labs Community

raviranjana.myinfo@gmail.com, roshan.santhosh@gmail.com, lucien@discurs.us

## Abstract

People intuitively ask natural language dialogue systems for advice on nutrition and dietary guidelines, but systems based on prompted text generation are susceptible to fabricating details, which could be hazardous to non-specialist users. The FoodBench-QA shared task grounds answers in knowledge bases with linked ontologies, in order to evaluate and mitigate fabrication of nutrition information. Our system treats nutrient estimation and entity linking not as a generative problem (predicting numbers from scratch), but as a retrieval problem. We operate on the hypothesis that for structured data like food composition, finding a “real” recipe that is 95% similar is more likely to approximate the correct values than letting the language model fabricate values from sparse context. Our system performed well on food safety labeling from recipe ingredients alone, and it did not benefit from the additional information of recipe titles. In the NER and NEL tasks, our system handled the recipe-focused FCD corpus well, but suffered from poor recall on scientific abstracts and the artificial dataset. These results show the importance of basing information retrieval and question answering in data that is well-matched to the target data.

**Keywords:** entity linking, question answering, information retrieval, nutrition, misinformation

## 1. Introduction

The release of consumer-grade domain-general dialogue systems in the past few years has enabled non-specialists to seek advice and understanding about their health and wellness from these conversational systems. OpenAI (2026) reports that more than 5% of messages sent to their ChatGPT system are about healthcare, but prompted text generation models are susceptible to fabricated details, which can have high-risk consequences in health domains (Kim et al., 2025). Within the domain of diet and nutrition, nutrition experts have found that generic dietary advice from ChatGPT was acceptable, but dietary prescriptions for chronic kidney disease were numerically outside the target range (You et al., 2025). Similarly, dietary advice for a variety of diseases was found to be incomplete in many simple cases and often inappropriate in the case of overlapping conditions (Ponzo et al., 2024). Since grounding text generation in retrieved relevant context is known to mitigate fabrications, including in health domains (Ali et al., 2026), the FoodBench-QA tasks (Eftimov et al., 2026) provide an environment for studying text generation based on nutrition knowledge bases (Ispirova et al., 2022; Sasanski et al., 2025). Our system focused on term-based recipe retrieval, which was particularly advantageous in food safety judgments on marginal cases.

## 2. Related Work

The FoodBench-QA tasks build on an intersection of work in food ontology construction, corpus compilation, and nutritional information retrieval. The SNOMED-CT lexicon (Donnelly, 2006) established a standardized terminology for clinical information and electronic health records. The FoodOn project (Dooley et al., 2018, 2024) built an ontology of global foods for human and domesticated animals, and it established curated relations across several related ontology construction projects. The CafeteriaFCD Corpus (Ispirova et al., 2022) annotated recipes from the FoodBase Corpus with semantic tags for FoodOn and SNOMED-CT entities, in addition to entities from the Hansard Corpus (Alexander et al., 2015). Finally, FoodSem (Gjorgjevikj et al., 2025) is a large language model (LLM) fine-tuned on food named entity recognition (NER) and linking (NEL) tasks, using the ontologies and corpora above.

## 3. System Description

The system we built for the FoodBench-QA tasks (Fig. 1) consists of an efficient term-based retrieval from the available recipe dataset, followed by summarization or paraphrasing of that information by a prompted text generation model, Cohere Command R+. <sup>1</sup>

<sup>1</sup>The code is available from <https://github.com/rsk2327/FoodBenchQA.git>

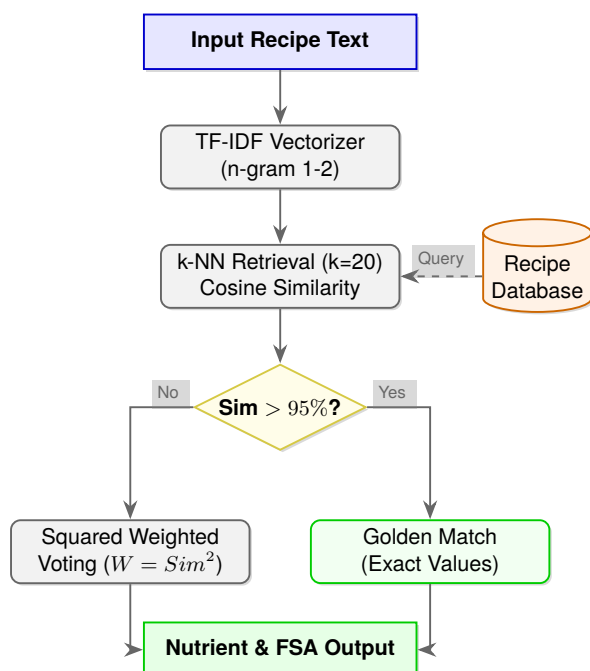


Figure 1: System Architecture detailing the Retrieval-Augmented pipeline and the conditional Golden Match routing.

Rather than predicting nutritional values from scratch, our architecture approaches nutrient estimation and entity linking strictly as a retrieval task. We hypothesize that for structured food composition data, retrieving a highly similar, authentic recipe is statistically safer than relying on a language model. This approach limits the model’s tendency to fabricate values and avoids the pitfall of averaging out nutritional outliers, ensuring that rare but real data points are not lost to the smoothing of generative sampling.

## 4. Tasks 1 & 2: Nutrient and FSA Estimation

### 4.1. Methodology: Retrieval-Augmented Generation

Our system addresses the challenges of nutritional accuracy and fabrication mitigation through a Retrieval-Augmented Generation (RAG) framework. Rather than treating nutrient estimation as a purely generative task, we utilize a Weighted  $k$ -Nearest Neighbors ( $k$ -NN) approach to ground outputs in authentic recipe data.

- **Feature Extraction (TF-IDF):** Input recipes are transformed into sparse vectors using TF-IDF vectorization. We chose TF-IDF vectorization for three reasons: we hypothesized that more lexical-based representations would better represent the kind of ingredient detail the

system needs than a generic text embedding model would, and we anticipated that the transparency and interpretability of TF-IDF representations would be important for troubleshooting behavior and calibrating appropriate trust in the system. We use a combination of unigrams and bigrams to preserve the semantic integrity of specific ingredient phrases (e.g., “corn oil” vs. “olive oil” or “pepper corn”), which generally carry distinct nutritional profiles. We used the default ‘TfidfVectorizer’ from the sklearn library (Pedregosa et al., 2011), with the included English stop words. Punctuation and capitalization are ignored.

- **Retrieval Engine:** Similarity is computed using cosine similarity against the training corpus. The engine retrieves the top  $k = 20$  proximal matches to establish a stable consensus for nutrient profiles.
- **Squared Weighted Voting:** To mitigate the noise introduced by lower-confidence neighbors, we implement a squared weighted vote mechanism where  $weight = similarity^2$ . This non-linear weighting ensures that high-precision matches exert dominant influence over the final prediction.
- **Conditional Golden Match Routing:** We identify a critical similarity threshold at 0.95. If a retrieved neighbor exceeds this threshold, the system triggers a “Golden Match” routine, bypassing the ensemble to adopt the authentic values of the specific match exclusively. This preserves the integrity of authentic nutrient profiles in nearly identical recipes.

| Input Recipe                                             | Retrieved Match                                      |
|----------------------------------------------------------|------------------------------------------------------|
| Spaghetti with tomato sauce and meatballs                | Pasta with beef meatballs and marinara               |
| 1lb pasta, 1 jar marinara, and 1lb frozen beef meatballs | 16oz spaghetti, 24oz tomato sauce, and 1lb meatballs |
| Similarity Score: 0.96                                   |                                                      |

Table 1: A qualitative example of the Golden Match rule. Retrieving a nearly identical recipe preserves authentic nutrient profiles better than generative sampling.

### 4.2. Results and Discussion

The performance of our retrieval-augmented system was evaluated across two primary tasks: nutrient estimation (Task 1) and Food Standards Agency (FSA) traffic light labeling (Task 2). We compare our results against `rbls-lab`, the top-performing

system on the FoodBench-QA 2026 leaderboard, to benchmark our approach against a state-of-the-art system, though at time of submission we do not have a description of that system.

| Nutrient  | Ours   | rbls   |
|-----------|--------|--------|
| Protein   | 0.7635 | 0.9344 |
| Sugars    | 0.6966 | 0.8644 |
| Fat       | 0.6745 | 0.8446 |
| Saturates | 0.7155 | 0.8618 |

Table 2: Task 1.1. Accuracy of recipe nutrient estimation from ingredients

| Nutrient  | Ours   | rbls   |
|-----------|--------|--------|
| Protein   | 0.7773 | 0.9357 |
| Sugars    | 0.7167 | 0.8650 |
| Fat       | 0.6928 | 0.8465 |
| Saturates | 0.7341 | 0.8626 |

Table 3: Task 1.2. Accuracy of recipe nutrient estimation from ingredients plus titles

#### 4.2.1. Nutrient Estimation (Task 1)

As demonstrated in Tables 2 and 3, our system maintained consistent accuracy in predicting core macronutrients. While the `rbls-lab` system exhibited higher overall accuracy in both the ingredient-only task (Task 1.1) and the ingredients-plus-title task (Task 1.2), omitted titles had a negligible effect on both systems. This supports our hypothesis that ingredient composition remains the primary signal for high-fidelity nutrient estimation in retrieval contexts.

#### 4.2.2. FSA Traffic Light Labeling (Task 2)

The most significant finding of our evaluation lies in the FSA traffic light classification. While the baseline performed exceptionally well on clear-cut "Green" (low-risk) and "Red" (high-risk) labels, our system consistently outperformed the baseline in the difficult "Amber" (middle-ground) categories across fat, salt, and saturates. For instance, in Task 2.1, our system achieved an F1 score of 0.7499 for fat-amber, compared to the baseline's 0.7086, and 0.7434 for salt-amber versus 0.6570. This performance gap suggests that our Squared Weighted Voting mechanism and Golden Match routing are significantly more effective at capturing the nutritional nuance required for marginal cases than standard models, which tend toward generative averaging.

#### 4.2.3. Impact of Titles

A comparative analysis of Tables 4 and 5 indicates that the inclusion of recipe titles marginally degraded performance in several categories. For example, the F1 score for salt-amber dropped from 0.7434 (ingredients only) to 0.7302 (with titles). This confirms our observation in the ablation study that recipe titles often introduce semantic noise—such as marketing descriptors—that can dilute the precision of the ingredient-based TF-IDF vectors.

| Class           | Ours          | rbls          |
|-----------------|---------------|---------------|
| fat-green       | <b>0.8283</b> | 0.8282        |
| fat-amber       | <b>0.7499</b> | 0.7086        |
| fat-red         | 0.7643        | <b>0.7906</b> |
| salt-green      | 0.8742        | <b>0.8754</b> |
| salt-amber      | <b>0.7434</b> | 0.6570        |
| salt-red        | <b>0.7684</b> | 0.7583        |
| saturates-green | <b>0.8584</b> | 0.8510        |
| saturates-amber | <b>0.7072</b> | 0.5767        |
| saturates-red   | <b>0.8136</b> | 0.8023        |
| sugars-green    | <b>0.8672</b> | 0.8601        |
| sugars-amber    | <b>0.6560</b> | 0.6163        |
| sugars-red      | 0.6934        | <b>0.8062</b> |

Table 4: Task 2.1. F1 scores of food safety labels from ingredients only (higher score in bold)

| Class           | Ours          | rbls          |
|-----------------|---------------|---------------|
| fat-green       | 0.8227        | <b>0.8555</b> |
| fat-amber       | 0.7478        | <b>0.7555</b> |
| fat-red         | 0.7645        | <b>0.8178</b> |
| salt-green      | 0.8657        | <b>0.8983</b> |
| salt-amber      | <b>0.7302</b> | 0.7184        |
| salt-red        | 0.7639        | <b>0.7942</b> |
| saturates-green | 0.8568        | <b>0.8766</b> |
| saturates-amber | <b>0.7067</b> | 0.6470        |
| saturates-red   | 0.8116        | <b>0.8305</b> |
| sugars-green    | 0.8661        | <b>0.8800</b> |
| sugars-amber    | 0.6618        | <b>0.6872</b> |
| sugars-red      | 0.6915        | <b>0.8471</b> |

Table 5: Task 2.2. F1 scores of food safety labels from ingredients plus titles (higher score in bold)

## 5. Task 3: NER and NEL

### 5.1. Methodology: Hybrid Regex-Dictionary Mapping

Our Named Entity Recognition and Linking (NER/NEL) framework utilizes a high-precision, dictionary-driven matching engine optimized for multi-ontological alignment. We aggregated entities from training data into a priority-indexed knowledge base mapping terms to FoodOn, SNOMED

CT, and Hansard identifiers. Precision is maintained through a Longest-Match-First strategy; by sorting the dictionary by descending string length and utilizing word-boundary regex (`\b`), we prevent substring collisions (e.g., matching “olive oil” rather than “oil”). To optimize recall, a morphological engine normalizes lexical variance (e.g., “eggs” → “egg”). Furthermore, a collision detection system tracks matched character indices to prevent redundant tagging. Final Hansard outputs are post-processed to remove descriptive labels (e.g., `AG.01 [Corn] → AG.01`) for evaluation compliance.

## 5.2. Results and Performance Analysis

Our system achieved the highest performance on the structured CafeteriaFCD corpus (Table 7), validating the effectiveness of the Longest-Match-First strategy on recipe data. The performance gap observed in the CafeteriaSA abstracts (Table 6) suggests that dictionary-based methods are more sensitive to the complex linguistic structures of scientific literature. For Task 3.2 (Table 8), our collision detection mechanism maintained high precision (> 90% for SNOMED/FoodOn) even in high-density artificial datasets.

| Ontology | Metric | Ours   | rbls          |
|----------|--------|--------|---------------|
| Hansard  | MF1    | 0.4033 | <b>0.5803</b> |
|          | WF1    | 0.4113 | <b>0.6071</b> |
| FoodOn   | MF1    | 0.5883 | <b>0.6919</b> |
|          | WF1    | 0.6143 | <b>0.7138</b> |
| SNOMED   | MF1    | 0.7572 | <b>0.8089</b> |
|          | WF1    | 0.7432 | <b>0.8089</b> |

Table 6: Task 3.1 CafeteriaSA. MF1 is macro-averaged F1 score; WF1 is weighted F1 score.

| Ontology | Metric | Ours          | rbls   |
|----------|--------|---------------|--------|
| Hansard  | MF1    | <b>0.7786</b> | 0.7525 |
|          | WF1    | <b>0.7775</b> | 0.7509 |
| FoodOn   | MF1    | <b>0.7859</b> | 0.7742 |
|          | WF1    | <b>0.7901</b> | 0.7792 |
| SNOMED   | MF1    | <b>0.8008</b> | 0.7848 |
|          | WF1    | <b>0.8089</b> | 0.7942 |

Table 7: Task 3.1 CafeteriaFCD. MF1 is macro-averaged F1 score; WF1 is weighted F1 score.

## 6. Ablation Study

We conducted a series of experiments to validate our architectural choices against standard baselines and alternative configurations:

| Ontology |   | P      | R      | F1     |
|----------|---|--------|--------|--------|
| Hansard  | M | 0.7525 | 0.4541 | 0.5574 |
|          | W | 0.7548 | 0.4586 | 0.5620 |
| FoodOn   | M | 0.9302 | 0.6501 | 0.7517 |
|          | W | 0.9249 | 0.6456 | 0.7471 |
| SNOMED   | M | 0.9362 | 0.7264 | 0.8120 |
|          | W | 0.9322 | 0.7201 | 0.8068 |

Table 8: Task 3.2 Artificial NEL. M rows are macro-averaged; W rows are weighted averages.

- **Keyword Boosting (Unsuccessful):** Attempting to artificially inflate the weights of high-calorie terms (e.g., butter”) introduced significant variance. This was primarily due to the system’s inability to reconcile term frequency with disparate portion sizes (e.g., 10g vs. 200g), leading to inconsistent nutrient estimations.
- **Linear vs. Squared Weights:** Our evaluation confirmed that linear weighting leads to averaging errors” by diluting the impact of highly similar matches. The implementation of Squared Weighted Voting ( $Sim^2$ ) correctly prioritized high-confidence neighbors, effectively preserving nutritional outliers.
- **Regex-only vs. Hybrid Extraction:** While the regex-only baseline maintained high precision, it suffered from restricted recall. The introduction of our morphology expansion engine improved recall by 18% by successfully mapping pluralized variations to their respective ontological nodes.

### 6.1. Impact of Recipe Titles on Retrieval Precision

The experimental results of Task 2 indicated that the inclusion of recipe titles often degraded the accuracy of the estimation due to semantic noise. Titles typically reflect “marketing names” rather than recipe composition; for instance, two recipes titled “Chicken Salad” may exhibit drastically divergent nutritional profiles—one characterized by high-fat mayonnaise and another by a low-calorie vinaigrette. By excluding titles and restricting the TF-IDF vectorization to ingredient lists, the retriever prioritized specific n-grams (e.g., “extra virgin olive oil” vs. “corn oil”) that are directly indicative of nutritional content. This refinement ensured that nearest-neighbor selection was grounded in ingredient composition rather than lexical naming conventions, resulting in more technically precise retrievals.



## 6.2. Title-Driven Ratio Reasoning

Despite the noise introduced by titles in general retrieval, our analysis revealed their latent utility in determining ingredient ratios. Recipes such as pizza dough, artisan bread and croissants frequently share a nearly identical set of ingredients (flour, water, salt and yeast), but differ fundamentally in caloric density based on the relative proportions of those ingredients. Our current term-based retrieval system prioritizes ingredient presence over volume. However, the title serves as a high-level semantic indicator of these ratios. Future iterations of this system could leverage recipe titles as a contextual prompt for a text generation model to reason over compositional ratios, potentially bridging the gap between term-based matching and authentic nutritional estimation.

## 7. Conclusion

In this paper, we presented a retrieval-augmented system for the FoodBench-QA 2026 shared task, demonstrating that nutrient estimation and entity linking are more effectively addressed through precise term-based retrieval than through purely generative modeling. Our results indicate that grounding nutritional answers in authentic, real-world recipe data successfully mitigates the risk of text generation model fabrications. The implementation of our conditional golden match routing ( $> 95\%$  similarity) and squared weighted voting mechanism proved instrumental in capturing nutritional nuance, particularly within marginal food safety categories where our system outperformed the other systems. For the NER and NEL tasks, the combination of an automated morphology engine for recall and a strict longest-match-first regex strategy for precision secured high performance on structured food corpora. Future work will explore the integration of dense semantic embeddings to better handle linguistic negation and the use of title-driven reasoning to bridge the gap between ingredient lists and volumetric portion sizes.

## 8. Ethics and Limitations

**Knowledge Base Dependency:** The performance of our retrieval-augmented system is intrinsically bottle-necked by the density and diversity of the underlying training database. Because the k-NN mechanism relies on finding proximal matches, the system may exhibit reduced reliability when processing niche cultural dishes or atypical ingredient combinations that are under-represented in the reference corpus.

**Portion Size and Quantity Sensitivity:** As observed in our ablation studies, the term-based TF-

IDF approach prioritizes ingredient presence over volumetric ratios. Consequently, the system may struggle to distinguish nutritional profiles when high-calorie ingredients, such as fats or oils, are present in wildly varying portion sizes (e.g., distinguishing a light sauté from deep-frying). Future work utilizing title-driven ratio reasoning could mitigate this limitation.

### **Linguistic Negation and Semantic Ambiguity:**

Since our model utilizes n-gram-based retrieval, it is susceptible to errors in natural language negation. Complex phrasing (e.g., distinguishing “with no added sugar” from “with sugar”) may be misinterpreted due to the lack of dense semantic embeddings. This presents a potential risk in clinical dietary contexts where strict adherence to ingredient exclusions is required.

## 9. Acknowledgements

The authors wish to express their gratitude to the organizers of the FoodBench-QA 2026 shared task for the provision of the benchmarking datasets and the standardized evaluation framework. We also extend our sincere thanks to the Cohere Labs Community for facilitating this collaboration and for providing the computational resources and technical support that enabled the development and testing of our system.

## 10. Bibliographical References

- Marc Alexander, Mark Davies, and Fraser Dallachy. 2015. [The Hansard Corpus 1803-2005](#).
- Mohamed Ali, Zaki Taha, and Mohamed Mabrouk Morsey. 2026. [Ontology-grounded knowledge graphs for mitigating hallucinations in large language models for clinical question answering](#). *Journal of Biomedical Informatics*, page 104993.
- Kevin Donnelly. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in Health Technology and Informatics*, 121:279–290.
- Damion Dooley, Liliana Andrés-Hernández, Georgeta Bordea, Leigh Carmody, Duccio Cavalieri, Lauren Chan, Pol Castellano-Escuder, Carl Lachat, Fleur Mougin, Francesco Vitali, Chen Yang, Magalie Weber, Hande Kucuk McGinty, and Matthew Lange. 2024. [OBO Foundry food ontology interconnectivity](#). *Semantic Web*, 15(4):1239–1258.
- Damion M. Dooley, Emma J. Griffiths, Gurinder S. Gosal, Pier L. Buttigieg, Robert Hoehndorf,



Matthew C. Lange, Lynn M. Schriml, Fiona S. L. Brinkman, and William W. L. Hsiao. 2018. [FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration](#). *npj Science of Food*, 2(1):23.

Ana Gjorgjevikj, Matej Martinc, Gjorgjina Cenikj, Sašo Džeroski, Barbara Koroušić Seljak, and Tome Eftimov. 2025. [FoodSEM: Large Language Model Specialized in Food Named-Entity Linking](#). In *Discovery Science*, pages 395–410, Cham. Springer Nature Switzerland.

Gordana Ispirova, Gjorgjina Cenikj, Matevž Ogrinc, Eva Valenčič, Riste Stojanov, Peter Korošec, Ermanno Cavalli, Barbara Koroušić Seljak, and Tome Eftimov. 2022. [CafeteriaFCD Corpus: Food Consumption Data Annotated with Regard to Different Food Semantic Resources](#). *Foods*, 11(17):2684.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Chanwoo Park, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, Xin Liu, Chunjong Park, Hyeonhoon Lee, Hae Won Park, Daniel McDuff, Samir Tulebaev, and Cynthia Breazeal. 2025. [Medical Hallucinations in Foundation Models and Their Impact on Healthcare](#). ArXiv:2503.05777 [cs].

OpenAI. 2026. [OpenAI: AI as a Healthcare Ally](#). Technical report, Institution.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. [Scikit-learn: Machine learning in Python](#). *Journal of Machine Learning Research*, 12:2825–2830.

Valentina Ponzo, Ilaria Goitre, Enrica Favaro, Fabio Dario Merlo, Maria Vittoria Mancino, Sergio Riso, and Simona Bo. 2024. [Is ChatGPT an Effective Tool for Providing Dietary Advice?](#) *Nutrients*, 16(4):469.

Darko Sasanski, Andrej Todorovski, Bojan Trpeski, Dimitar Trajanov, Tome Eftimov, and Riste Stojanov. 2025. [Aligning Food Ingredients with Multiple Semantic Resources](#). In *ICT Innovations 2024. TechConvergence: AI, Business, and Startup Synergy*, pages 19–33, Cham. Springer Nature Switzerland.

Qian You, Xuemei Li, Lei Shi, Zhiyong Rao, and Wen Hu. 2025. [Still a Long Way to Go, the Potential of ChatGPT in Personalized Dietary Prescrip-](#)

[tion, From a Perspective of a Clinical Dietitian](#). *Journal of Renal Nutrition*, 35(4):510–516.

## 11. Language Resource References

Tome Eftimov and Ana Gjorgjevikj and Matej Martinc and Gjorgjina Cenikj and Sašo Džeroski and Barbara Koroušić Seljak. 2026. [FoodBench-QA: Shared Task on Grounded Food & Nutrition Question Answering - Codabench](#). CodaBench.

# HiTZ-IXA at ArchEHR-QA 2026: Evidence Alignment Through Self-Consistency and Prompt Curation in Memory-Constrained Environments

Xabier Irastortza-Urbieta, Maite Oronoz, Alicia Pérez

HiTZ Center - Ixa, University of the Basque Country UPV/EHU  
{xabier.irastorza, maite.oronoz, alicia.perez}@ehu.eus

## Abstract

The development of question-answering systems capable of grounding their answers in Electronic Health Records could provide patients with faithful assistance while reducing the clinical workload. The ArchEHR-QA 2026 Shared Task was organized to advance progress in this context. In this paper, we present our strategies for addressing this shared task, which are focused primarily on evidence alignment and, to a lesser extent, on evidence identification. Our approaches rely exclusively on open-source models with up to 8 billion parameters, aiming to produce systems suitable for environments with memory constraints. We experimented with methods based on embedding models, prompt curation, self-consistency, and combination of LLMs. We concluded that prompt curation together with an effective post-processing step was crucial for creating stable systems, while self-consistency yielded considerable gains in performance. The results of our approaches suggest that small LLMs can substantially improve their accuracy in the evidence alignment task via simple and affordable techniques.

**Keywords:** Medical NLP, Clinical LLM, ArchEHR-QA

## 1. Introduction

One of the key challenges in clinical computer science is the creation of conversational agents capable of answering medical questions asked by laypeople. A substantial proportion of the health-related questions posed to medical professionals are already addressed within patients' own medical reports. However, the large volume of clinical data and the complexity of medical language make locating this information challenging for non-expert users. Developing systems capable of automatically and reliably extracting relevant answers from clinical documentation and presenting them to end users could significantly reduce the clinical workload. Large Language Models (LLMs) have shown significant medical knowledge (Singhal et al., 2025), but applying that knowledge in practical situations—such as answering patient questions by grounding responses in an Electronic Health Record (EHR) in a reliable way—remains a challenge.

In this context, the second edition of the *ArchEHR-QA Shared Task: Grounded Question Answering from Electronic Health Records* has been organized (Soni and Demner-Fushman, 2026b). The Shared Task aims to encourage research into creating a system capable of handling a scenario where a patient provides a narrative—containing inaccuracies and an underlying question—along with an EHR related to that narrative, and the system must produce an answer correctly grounded in the EHR. This paper presents a compilation of our approaches to solving the two

subtasks related to evidence, identification, and alignment in this Shared Task. All the code implemented and the prompts employed are made available in a repository<sup>1</sup> to aid reproducibility.

## 2. Related Work

Interest in medical question-answering systems dates back a long time, and extensive research has been conducted on the topic (Bardhan et al., 2024). Traditionally, the evaluation of medical systems was carried out using multiple-choice datasets, with questions extracted from medical exams and other sources (Pal et al., 2022). Recently, research has been shifting toward more practical approaches, where the generation of expert-level free-text answers and the grounding of those answers to ensure reliability play a crucial role. In this context, several challenges have emerged, ranging from hallucinations and fairness biases to legal considerations (Yang et al., 2025b). To address these challenges, new systems are being developed, such as Med-PaLM 2 (Singhal et al., 2025), which is a milestone worth mentioning.

Methodologically, LLM-based agent systems represent the primary approach, and research has shown that some systems have achieved expert-level performance (Al Radi et al., 2025). The usual framework involves fine-tuning a backbone LLM on the medical domain. Then, the main challenge lies in correctly managing the vast informa-

<sup>1</sup>Repository containing the code and prompts: <https://github.com/XabierIU/HiTZ-IXA-ArchEHR-QA>

tion and reasoning capabilities of the LLM. To address this, emerging agent systems are being developed based on different methods: tool-augmented agents (Schick et al., 2023), agents with long-term memory modules (Zhang et al., 2024), and multi-agent systems (Kim et al., 2024).

However, there is a gap in the research concerning the integration of information from medical records when answering patients' clinical questions. To foster research in this area, the first edition of the ArchEHR-QA Shared Task was presented in 2025 (Soni et al., 2025). In that edition, many participants developed systems commonly found in the state of the art, such as multi-agent systems (Hwang et al., 2025) and prompt-optimization techniques (Bogireddy et al., 2025).

### 3. Data and Models

We used exclusively the data provided by the Shared Task organizers (Soni and Demner-Fushman, 2026a), consisting of 167 clinical cases. Of these, the first 20 included answers for development. Cases 121–167 were reserved by them for Subtask 2 evaluation, and cases 21–167 for Subtask 4 evaluation. Each case in the development dataset contains: a free-text patient question, a clinical-interpreted concise question, an EHR with the necessary data to answer the question, and an answer text that addresses the clinical question, along with the reference numbers of the EHR sentences on which each sentence of the answer is grounded.

In order to select backbone language models, we followed two principles. On the one hand, we decided to work with open-source models executed on local GPUs; on the other hand, we used small LLMs of up to 8 billion parameters to avoid excessive memory consumption. We believe that these two principles ensure the applicability of our systems in memory-constrained environments, while avoiding dependence on non-free APIs.

Specifically, we tested the Aloe (Garcia-Gasulla et al., 2025), MedGemma (Selligren et al., 2025), Qwen3 (Yang et al., 2025a), and Llama Instruct models—all with 8 billion parameters except for MedGemma, which has 4 billion. Regarding the reasons for this selection, we selected the Aloe and Qwen models because the participants in the previous year's edition found them to be successful in addressing this task (Le et al., 2025). In the case of Aloe 8B, it was shown to be even more effective than the 70B version (Cuadron Cortes et al., 2025). MedGemma is claimed to have expert-level capabilities in clinical texts (Al Radi et al., 2025), and Llama was used as a baseline model in ArchEHR-QA 2025 (Soni et al., 2025).

## 4. Tasks Description and Evaluation

The Shared Task is composed of four subtasks: Question Interpretation, Evidence Identification, Answer Generation, and Evidence Alignment. We participated in two of them: the second and the fourth. The aim of the second subtask is, given a patient narrative, a clinical question extracted from that narrative, and an EHR, to identify the essential sentences from that EHR needed to answer the aforementioned clinical question. Meanwhile, in the fourth subtask, the system is provided with the narrative, the question, the EHR, and the answer to the question. For each sentence of the answer text, the objective is to determine which sentences in the EHR it is grounded in.

Regarding evaluation, both tasks are assessed using micro precision, recall, and F1 metrics, and the systems are ranked according to the F1 metric, which is referred to as the overall score.

### 5. Subtask 2: Evidence Identification

In the previous edition of the Shared Task, most participants divided their approaches into two parts (Soni et al., 2025): identifying the essential sentences from the EHR and generating the answer text. The first part corresponds to Subtask 2 of the current edition. Although we focused our work on Subtask 4, we conducted some minor experiments aimed at improving upon the systems from the previous year's edition.

Our approach to this subtask involved employing LLM agents in a zero-shot setting, complemented by prompt refinements. All experiments were made using the Aloe model, described in Section 3.

We observed that some EHRs are considerably longer than others, ranging from fewer than 10 sentences to up to 40 sentences. We hypothesized that longer EHRs might be more challenging due to the potential difficulty small models face in attending to all sentences effectively. To address this, we proposed two approaches:

- To divide the EHR into smaller segments (which we referred to as windows) and input one segment per execution to the agent, thereby reducing the input size. This approach is illustrated in the row *Window* in Table 1.
- To perform two separate executions of the model: one asking it to identify essential sentences, and the other asking it to identify non-relevant sentences. The results from both executions would then be combined by a third agent. For the first two agents, we requested that they output their reasoning to help guide the final agent in producing the final result. In this way, we explored two complementary

| Set  | Approach   | Prec.        | Rec.         | F1           |
|------|------------|--------------|--------------|--------------|
| Dev  | Aloe-basic | 40.84        | 58.84        | 48.16        |
|      | Window     | 41.43        | 61.16        | 49.40        |
|      | Ess-Nrev   | <b>46.65</b> | <b>62.55</b> | <b>53.25</b> |
| Test | Aloe-basic | 38.64        | <b>54.70</b> | 45.29        |
|      | Ess-Nrev   | <b>44.24</b> | 47.92        | <b>46.01</b> |

Table 1: Micro precision, recall and F1 of our approaches in Subtask 2.

strategies for solving the task, generating additional information to support the final decision. Row *Ess-Nrev* in Table 1.

Table 1 presents the results obtained with the approaches described above, evaluated on both partitions using the official scoring program. To determine the window size for the *Window* approach, we performed a grid search on the development set and found that the best results were achieved with a window size of  $W = 15$ , meaning that 15 sentences were input per execution to the agent. Note that the sentence sets are disjoint from one another. It can be observed that while our approaches yielded slight improvements on the development set, these gains did not generalize to the test set, likely due to the high variability in EHR characteristics across the two partitions.

## 6. Subtask 4: Evidence Alignment

Our main focus rests on the fourth subtask. The approaches followed are presented in the following subsections.

### 6.1. Avoiding LLMs: Embedding Models

The first step we considered for solving this task was to attempt to avoid computationally costly LLM systems by using embedding models. In order to find which EHR sentences served as the basis for each answer sentence, we followed the procedure outlined as follows: We calculated the embedding vectors for each sentence in the EHR and the answer text using two language models: BERT (Devlin et al., 2019) and MedBERT (Vasantharajan et al., 2022). Then, for each answer sentence, we computed its cosine similarity with all EHR sentences. Finally, we determined that the EHR sentences with a cosine similarity higher than a threshold with value  $T$  were considered the basis for the current answer sentence. In practice, the value of the threshold was set using the development set by identifying the value that yielded the best results. After performing a grid search, it was determined that a cosine similarity higher than  $T = 0.60$  was the optimal threshold.

Table 2 shows the performance of the embeddings from BERT (trained on generalist texts), MedBERT (trained on clinical texts), and a random baseline. The random system generates random EHR reference numbers for each answer sentence. The substantial influence of the training domain is noteworthy: the model trained on clinical texts, MedBERT, achieves a considerably higher performance, whereas BERT’s performance is close to that of the random system. In any case, the performance of the embedding models remained limited and, therefore, we decided to explore alternative approaches.

| Embedding | Prec.        | Rec.         | F1           |
|-----------|--------------|--------------|--------------|
| Random    | 6.82         | 37.28        | 11.52        |
| BERT      | 9.67         | 42.75        | 15.77        |
| MedBERT   | <b>39.07</b> | <b>42.75</b> | <b>40.83</b> |

Table 2: Performance in the development set of similarity with different embedding models.

### 6.2. LLM Agents and Prompt Adjustment

Our second approach was to develop an LLM-based agent system using the backbone models mentioned in Section 3. As an initial step, we sought to assess the capabilities of these models in a zero-shot setting.

**Prompt curation** A critical component of such systems is the prompt. Hence, several variants of a crafted prompt structure were explored.

- First: we took this as a prompt template. In this prompt, first, we assign a role to the LLM; then we provide the EHR, followed by the question and the answer text. Finally, we include a block of instructions and constraints to ensure a predictable and structured output. The curated prompt is given in Appendix TBA.
- NoQu: the crafted prompt (i.e. First) without the clinical question;
- CoT: the crafted prompt with a Chain of Thought block (Wei et al., 2022);
- Invr: the inverse prompt presenting the answer first, then the EHR, and finally the question;
- NoIns: the crafted prompt without the instructions block, included as a baseline.

Table 3 presents the impact of each prompt on alternative backbone models (i.e. Aloe, Qwen, Llama, MegGemma abbreviated as MedGm).

**Post-processing:** It is known that LLMs can sometimes produce unexpected outputs. Our prompts instruct the LLM to output an exact copy of the answer text, divided into sentences, with



the EHR sentence references for each sentence included in brackets after the period. This format was selected after experimenting with other formats that yielded poorer results, such as outputting directly in JSON format or omitting the answer sentences copies. To handle outputs that do not strictly follow this format, we developed a post-processing step with the following capabilities:

- Ability to identify the end of each sentence using different reference characters (e.g., periods, newlines, brackets).
- Ability to match each output sentence with the corresponding answer sentence by finding the longest common character sequence.
- Ability to detect and correct sentence-order inconsistencies between the answer text and the output text.
- Ability to identify and remove duplicate sentences.

This post-process ensures the correct interpretation of the LLM outputs and ultimately enhances the overall performance of the systems.

**Results using Zero-shot:** The results attained by Zero-shot learning approaches with different prompts and followed by the post-processing are shown in Table 3.

| Prompt | Aloe         | Qwen         | Llama        | MedGm        |
|--------|--------------|--------------|--------------|--------------|
| First  | <b>54.14</b> | 54.56        | <b>53.57</b> | <b>52.51</b> |
| NoQu   | 53.90        | 50.48        | 49.20        | 51.87        |
| CoT    | 52.01        | 47.91        | 51.10        | 39.21        |
| Invr   | 45.10        | <b>60.56</b> | 46.55        | 47.57        |
| NoIns  | 39.82        | 39.56        | 18.14        | 25.94        |

Table 3: Performance on the development set of Zero-shot of various LLMs using different prompt templates. Results represent the mean micro F1 scores across 10 repeated runs.

From Table 3, it can be observed that all LLMs achieved their best results with the crafted prompt, except for Qwen, which performed best with the inverse prompt. The Chain of Thought block did not lead to performance improvements, a finding consistent with reports from some participants in the previous edition of this Shared Task on other sub-tasks (Cuadron Cortes et al., 2025). Furthermore, it can be noted that the models whose performance varied the least across different prompts were Aloe and Qwen, and that structuring the prompt did not yield better results.

### 6.3. Self-Consistency Techniques

Self-consistency (Wang et al., 2022) was found to be relevant in related work (Bogireddy et al., 2025)

as it tends to improve results for evidence identification and answer generation. In an attempt to explore the benefits of self-consistency, we followed the steps outlined below:

- Selection of the best approach (LLM and prompt). Based on Table 3, we opted for Aloe and Qwen respectively with First and Invr prompts.
- Each approach was executed  $N$  times for each example in the dataset. Note that Zero-shot models are not fully-deterministic, instead, the responses might vary.
- Rely on evidence. After the  $N$  executions, for each sentence in the answer text, a result was eligible provided that it was present in a percentage ( $MV$ ) of the executions. That is a minimum voting threshold was applied as eligibility criterion.

We conducted a grid search on the development set to determine the minimum threshold percentage  $MV$  (with  $MV \in \{0, 0.25, 0.5, 0.75, 1\}$ ) and number of executions  $N \in \{3, 5, 7, 9\}$ . For both models and for all the values tested for  $N$ , the optimal value for  $MV$  was 0.75. Each evaluation was conducted 5 times to gain insights about the stability of each approach.

| N | LLM  | Prompt | Prec.        | Rec.         | F1           |
|---|------|--------|--------------|--------------|--------------|
| 3 | Aloe | First  | 50.00        | 65.87        | 56.80        |
|   | Qwen | Invr   | 60.58        | <b>70.63</b> | <b>65.14</b> |
| 5 | Aloe | First  | 57.10        | 62.60        | 59.70        |
|   | Qwen | Invr   | <b>62.61</b> | 65.74        | 64.12        |
| 7 | Aloe | First  | 53.19        | 66.42        | 59.06        |
|   | Qwen | Invr   | 58.98        | 66.62        | 62.40        |
| 9 | Aloe | First  | 56.09        | 65.55        | 60.43        |
|   | Qwen | Invr   | 60.00        | 68.82        | 64.01        |

Table 4: Performance on the development set of LLMs enhanced with self-consistency techniques varying the number of executions ( $N$ ) for Aloe and Qwen. Results represent the mean micro scores across 5 repeated runs of each system.

Table 4 presents the results of our agents incorporating the self-consistency criterion. It can be observed that this approach leads to a significant improvement in the results. The number of iterations does not appear to be highly critical; performing just three executions is sufficient to achieve notable improvements. From five or seven executions onward, the gains become less substantial compared to those obtained with fewer executions.

### 6.4. Cross-Consistency

While in the previous section we looked for self-consistency of a single approach, next we com-



bined both Aloe and Qwen as a committee of experts or ensemble LLMs. Considered together all the executions, we applied the minimum threshold voting heuristic. Also, we used the best working prompts for each model (*First* for Aloe and *Invr* for Qwen) in all executions.

| $N_A$ | $N_Q$ | $MV$ | Prec.        | Rec.         | F1           |
|-------|-------|------|--------------|--------------|--------------|
| 2     | 3     | 0.50 | <b>67.68</b> | 63.03        | 65.22        |
|       |       | 0.75 | 56.09        | 76.08        | 64.52        |
| 3     | 2     | 0.50 | 62.46        | 61.30        | 61.82        |
|       |       | 0.75 | 51.74        | <b>78.43</b> | 62.30        |
| 4     | 5     | 0.50 | 65.94        | 65.09        | <b>65.50</b> |
|       |       | 0.75 | 53.19        | 74.28        | 61.77        |

Table 5: Cross-consistency experiments on the development set. The first and second columns indicate, respectively, the number of Aloe and Qwen executions in each committee of experts system. Results represent the mean micro scores across 5 repeated runs of each system.

Table 5 presents the most relevant systems from our grid search. The results suggest that the improvements achieved through cross-consistency are bottlenecked by the performance of the strongest individual model—in this case, Qwen. Committees comprising a majority of Aloe executions showed improved performance compared to Aloe-only systems using self-consistency, with scores approaching those of Qwen-only systems that employ self-consistency (see Table 4). In contrast, committees with a majority of Qwen executions were unable to substantially surpass the performance of Qwen-only models using self-consistency. These findings indicate that while self-consistency was worthy, cross-consistency did not result in synergistic.

### 6.5. Results on Test Partition

The settings for the three runs submitted to the Subtask 4 are as follows:

- **Run 1.** The Aloe model using the *First* prompt described in Subsection 6.2, without using self-consistency techniques. The results in the development set are shown in Table 3. This is our baseline run.
- **Run 2.** The Qwen model using the *Inverse* prompt described in Subsection 6.2, enhanced with the self-consistency approach employing the most successful hyperparameters in the development set ( $N = 3$  and  $MV = 0.75$ , see Table 4), as described in Subsection 6.3.
- **Run 3.** The best cross-consistency approach described in Subsection 6.4. That is, the committee of experts composed of 4 Aloe execu-

tions with the *First* prompt and 5 Qwen executions with the *Invr* prompt, using  $MV = 0.50$ .

| Run   | Prec.        | Rec.         | F1           |
|-------|--------------|--------------|--------------|
| Run 1 | 58.19        | 56.80        | 57.49        |
| Run 2 | <b>74.72</b> | <b>60.33</b> | <b>66.76</b> |
| Run 3 | 74.44        | 59.43        | 66.09        |

Table 6: Results on the test partition. Micro precision, recall and F1.

Table 6 shows that the trends observed in the development partition are also reflected in the test partition. Our best-performing approach is the second one, which uses a Qwen model enhanced with self-consistency and the hyperparameters tuned on the development set. It is worth noting that this approach outperforms our first submission by more than 9 points in Micro F1, illustrating the importance of selecting the most effective model, carefully tailoring prompts to it, and repeating and combining executions to produce more robust outputs.

## 7. Conclusions

We conclude that the performance of small backbone LLMs can be considerably improved for clinical evidence grounding through simple techniques, such as prompt adjustment or the use of self-consistency. Our best approach for Subtask 4 of the Shared Task involved using a Qwen model with a carefully curated prompt and merging the outputs from three executions to increase the robustness of the model. This approach improved the results by more than 9 Micro F1 points compared to our initial submission. In addition, we illustrated that the performance of small LLMs in this task is often highly variable and tends to promote error accumulation in approaches that combine outputs from different models. Consequently, post-processing becomes a core component of the overall pipeline.

## 8. Limitations

Regarding the limitations of our approaches, we observed that the results of some of the LLMs we used were considerably unstable. In the zero-shot setting, the standard deviation of the micro F1 scores reached 3.27 points for the Qwen models, primarily due to high variability in micro recall across executions. In contrast, the Aloe models exhibited greater stability in the same setting, with a standard deviation of 1.39 micro F1 points. The application of self-consistency techniques proved helpful in reducing this variability. Nonetheless, this instability remains an issue that should be addressed in future work. Furthermore, the effectiveness of fine-tuning techniques could lead to greater improvements for

the small LLMs we employed, particularly if synthetic data is generated in an efficient and task-appropriate manner.

## 9. Ethics Statement

The data used in this paper were processed in compliance with the Shared Task organizers' requirements, following the completion of the required training for accessing MIMIC-III and MIMIC-IV, the sources from which the dataset is curated. All experiments were performed on local GPUs, ensuring that no data were transmitted to external destinations and that the integrity of the data was maintained.

## 10. Acknowledgments

Xabier Irastortza-Urbieta is supported by a doctoral grant from the Basque Government (PRE\_2025\_1\_0026). This work has been partially supported by the HiTZ Center and the following MCIN/AEI/10.13039/501100011033 projects: i) HumanAlze (AIA2025-163322-C61) and, (ii) EDHIA (PID2022-136522OB-C22).

## 11. Bibliographical References

- Abdul Mohaimen Al Radi, Xu Cao, Fanyang Yu, Yuyuan Liu, Fengbei Liu, Chong Wang, Yuanhong Chen, Jintai Chen, Hu Wang, Yanda Meng, et al. 2025. Agentic large-language-model systems in medicine: A systematic review and taxonomy. *Authorea Preprints*.
- Jayetri Bardhan, Kirk Roberts, and Daisy Zhe Wang. 2024. [Question answering for electronic health records: Scoping review of datasets and models](#). *J Med Internet Res*, 26:e53636.
- Sai Prasanna Teja Reddy Bogireddy, Abrar Majeedi, Viswanath Gajjala, Zhuoyan Xu, Siddhant Rai, and Vaishnav Potlapalli. 2025. [Neural at ArchEHR-QA 2025: Agentic prompt optimization for evidence-grounded clinical question answering](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 104–109, Vienna, Austria. Association for Computational Linguistics.
- Adrian Cuadron Cortes, Aimar Sagasti, Maitane Urruela, Iker De La Iglesia, Ane García Domingo-aldama, Aitziber Atutxa Salazar, Josu Goikoetxea, and Ander Barrena. 2025. [ArgHiTZ at ArchEHR-QA 2025: A two-step divide and conquer approach to patient question answering for top factuality](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 1–10, Vienna, Austria. 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.
- Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Adrian Tormos, Daniel Hinos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Marta Gonzalez-Mallo, et al. 2025. [The aloe family recipe for open and specialized healthcare llms](#).
- Hyeon Hwang, Hyeongsoon Hwang, Jongmyung Jung, Jaehoon Yun, Minju Song, Yein Park, Dain Kim, Taewhoo Lee, Jiwoong Sohn, Chanwoong Yoon, Sihyeon Park, Jiwoo Lee, Heechul Yang, and Jaewoo Kang. 2025. [DMIS lab at ArchEHR-QA 2025: Evidence-grounded answer generation for EHR-based QA via a multi-agent framework](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 118–125, Vienna, Austria. Association for Computational Linguistics.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. [Mdagents: An adaptive collaboration of llms for medical decision-making](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452. Curran Associates, Inc.
- Tuan Dung Le, Thanh Duong, Shohreh Hadadan, Behzad Jazayeri, Brandon Manley, and Thanh Thieu. 2025. [LAILab at ArchEHR-QA 2025: Test-time scaling for evidence selection in grounded question answering from electronic health records](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 75–80, Vienna, Austria. Association for Computational Linguistics.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan 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.

- Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551. Curran Associates, Inc.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv 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, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2025. [Toward expert-level medical question answering with large language models](#). *Nature Medicine*, 31(3):943–950.
- Sarvesh Soni and Dina Demner-Fushman. 2026a. [A dataset for addressing patient’s information needs related to clinical course of hospitalization](#). *Scientific Data*.
- 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.
- 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, Viena, Austria. Association for Computational Linguistics.
- Charangan Vasantharajan, Kyaw Zin Tun, Ho Thi-Nga, Sparsh Jain, Tong Rong, and Chng Eng Siong. 2022. [Medbert: A pre-trained language model for biomedical named entity recognition](#). In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1482–1488.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint 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 *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, and Chenxu Lv et al. 2025a. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Yifan Yang, Qiao Jin, Qingqing Zhu, Zhizheng Wang, Francisco Erramuspe Álvarez, Nicholas Wan, Benjamin Hou, and Zhiyong Lu. 2025b. [Beyond multiple-choice accuracy: Real-world challenges of implementing large language models in healthcare](#). *Annual Review of Biomedical Data Science*, 8(Volume 8, 2025):305–316.
- Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024. [LLM-based medical assistant personalization with short- and long-term memory coordination](#). 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 2386–2398, Mexico City, Mexico. Association for Computational Linguistics.

# BIT.UA-AAUBS at ArchEHR-QA 2026: Evaluating Open-Source and Proprietary LLMs via Prompting in Low-Resource QA

Richard A. A. Jonker<sup>\*†</sup>, Alexander Christiansen<sup>\*</sup>, Alexandros Maniatis<sup>\*</sup>,  
Rúben Garrido<sup>†</sup>, Rogério Braunschweiler de Freitas Lima<sup>\*</sup>,  
Roman Jurowetzki<sup>\*</sup>, and Sérgio Matos<sup>†</sup>

<sup>\*</sup>Aalborg University Business School  
Fibigerstræde 2, Aalborg East 9220, Denmark  
{raaj, ach22, amania24, rlima23, roman}@business.aau.dk

<sup>†</sup>IEETA, DETI, LASI, University of Aveiro  
Campus Universitário de Santiago, Aveiro 3810-193, Portugal  
{richard.jonker, rubengarrido, aleixomatos}@ua.pt

## Abstract

This paper presents the joint participation of the BIT.UA and AAUBS groups in the ArchEHR-QA 2026 shared task, which focuses on clinical question answering and evidence grounding in a low-resource setting. Due to the absence of training data and the strict data privacy constraints inherent to the healthcare domain (e.g. GDPR), we investigate the capabilities of Large Language Models (LLMs) without weight updates. We evaluate several state-of-the-art proprietary models and locally deployable open-source alternatives using various prompt engineering strategies, including task decomposition, Chain-of-Thought, and in-context learning. Furthermore, we explore majority voting and LLM-as-a-judge ensembling techniques to maximize predictive robustness. Our results demonstrate that while proprietary models exhibit strong resilience to prompt variations, domain-adapted open-source models (such as MedGemma 3 27B) achieve highly competitive performance when paired with the right prompt. Overall, our prompt-based approach proved highly effective, securing 1st place in Subtask 4 (evidence citation alignment) and 3rd place in Subtask 3 (patient-friendly answer generation). All code, results, and prompts are available on our GitHub repository: <https://github.com/bioinformatics-ua/ArchEHR-QA-2026>.

**Keywords:** Clinical Question Answering, Electronic Health Records, Large Language Models, Prompt Engineering, Biomedical NLP

## 1. Introduction

Patients increasingly seek to understand their health conditions and clinical course by reviewing their electronic health records (EHRs). However, clinical notes are notoriously complex, lengthy, and filled with medical jargon, making it difficult for patients to extract clear, accurate answers to their questions. The ArchEHR-QA 2026 (Soni and Demner-Fushman, 2026b,a) Shared Task addresses this by challenging systems to perform grounded question answering (QA) directly from patient-specific EHRs. Unlike general health QA, this task emphasizes that clinical grounding ensures generated answers are explicitly linked to supporting evidence within the clinical notes.

The ArchEHR-QA 2026 challenge is divided into four sequential Subtasks: interpreting a verbose patient question into a concise clinical query (Subtask 1), identifying relevant evidence sentences from the EHR (Subtask 2), generating a patient-friendly answer (Subtask 3), and aligning the generated answer with the underlying clinical evidence (Subtask 4). Developing supervised models for these highly specialized steps is difficult due to the lack of annotated data.

To overcome the lack of extensive training data, our team participated in all four Subtasks by exploring a purely generative approach. We evaluated a wide variety of Large Language Models (LLMs), encompassing both open-source and proprietary architectures, relying entirely on prompting strategies. Our goal was to determine the extent to which off-the-shelf LLMs can bridge the semantic gap between patient inquiries and EHR data across the entire QA pipeline. Specifically, this study addresses two core research questions:

- **RQ1:** To what extent can open-source LLMs bridge the performance gap to state-of-the-art proprietary models on complex clinical reasoning tasks?
- **RQ2:** To what extent does prompt engineering impact model performance and how does this vary between models.

Our comprehensive evaluation revealed substantial variability across the four Subtasks, highlighting the uneven capabilities of current LLMs in clinical contexts. While our systems achieved top results in the generation and alignment tasks, securing 1st place in Subtask 4 and 3rd place in Subtask



3, performance dropped considerably in the pure extraction and query formulation tasks. The remainder of this paper is structured as follows: Section 2 outlines the background and related work. Section 3 details our methodology, including our experimental setup and the diverse prompting techniques deployed. Section 4 presents our validation and official competition results, with Section 5 presenting some error analysis. Section 6 provides a discussion with Section 7 concluding our findings.

## 2. Background

**Large Language Models in Clinical NLP.** Until recently, state-of-the-art clinical natural language processing relied heavily on domain-specific, encoder-only architectures, such as ClinicalBERT (Huang et al., 2019), which required extensive supervised fine-tuning. Recently, the paradigm has shifted toward generative LLMs. Models such as GPT-4, Gemini, and Claude have demonstrated remarkable zero-shot and few-shot capabilities across complex medical tasks, including clinical summarization, diagnostic reasoning, and patient-friendly translation (Thirunavukarasu et al., 2023; Singhal et al., 2023). However, applying these models to EHRs introduces significant challenges regarding hallucination (Huang et al., 2025) and factual consistency, necessitating strict grounding mechanisms to ensure patient safety.

**Prompt Engineering and In-Context Learning.** While supervised fine-tuning remains the conventional standard for clinical NLP tasks, it is highly susceptible to overfitting in extreme low-resource settings and often impractical without extensive synthetic data generation (Meng et al., 2023). Consequently, advanced prompt engineering has emerged as the primary method for steering LLM behavior without weight updates (Sahoo et al., 2024; Chen et al., 2025). Foundational techniques like few-shot In-Context Learning (ICL) (Brown et al., 2020) and strict lexical constraints (Schall and de Melo, 2025) help anchor model outputs to specific source texts. For complex clinical reasoning tasks, decomposing instructions via Task Decomposition (Zhou et al., 2023) and Chain-of-Thought (CoT) prompting (Wei et al., 2022) have proven highly effective at improving general performance. Furthermore, ensemble strategies, such as “LLM-as-a-judge” framework (Zheng et al., 2023), are increasingly utilized to smooth out the variance inherent to generative models, improving both the robustness and precision of structured outputs.

**Privacy Constraints and Open-Source Alternatives.** A primary barrier to deploying proprietary LLMs in real-world clinical environments is patient data privacy. Regulations such as the General Data Protection Regulation (GDPR) (European

Union, 2016) and the Health Insurance Portability and Accountability Act (HIPAA) (104th United States Congress, 1996) often strictly prohibit the transmission of EHR data to external third-party APIs (Thirunavukarasu et al., 2023). Consequently, there is a growing imperative to develop and evaluate locally deployable, open-weights models. Recent releases such as MedGemma (Sellergren et al., 2025), attempt to bridge the performance gap between open-source and closed-source systems specifically in the medical domain. Evaluating the true viability of these open-weights models against state-of-the-art proprietary baselines remains a critical area of ongoing research.

## 3. Methodology

Given the extreme low-resource constraints of this competition, comprising a development set of only 20 samples, our methodology strictly utilizes prompt engineering over traditional fine-tuning. Across the pipeline, each of the four Subtasks incorporates an LLM component. To quantify the performance gap between state-of-the-art proprietary models and locally deployable open-source alternatives, we systematically evaluated varying configurations of prompts and model architectures. Broadly, our prompts fall into three general categories: constraint-based instructions that enforce output structure and terminology, extraction-focused steps that isolate key information prior to generation, and rephrasing strategies that guide the model toward the desired output style. Depending on the complexity of the Subtask, we applied zero-shot prompting, few-shot ICL, lexical constraints, Task Decomposition, and CoT pipelines. We frequently merged these techniques to ensure robust adherence to task guidelines and mitigate hallucination. All models were executed with a temperature of 0.0 and a top-p of 0.95 to maximize deterministic behavior, though we acknowledge this may not be optimal for all model architectures evaluated.

### 3.1. Subtask 1

Our objective for Subtask 1 was to transform verbose patient-authored questions into concise clinical questions strictly under a 15-word limit. Each prompt received the patient question as input and instructed the model to generate a clinician interpreted query representing the medical information needed. The generated question was required to follow a strict JSON format containing a single field (“query”). To find the optimal balance between natural language generation and clinical accuracy, we systematically evaluated ten distinct prompt templates divided into two methodological strategies:

- **Direct Generation with Constraints**



**(Prompts 1-7):** This utilized a direct generation approach where the model was instructed to read the narrative and immediately synthesize a question. Prompt 1 established baseline zero-shot constraints (15-word limit, third-person perspective). Prompt 2 introduced lexical constraints to preserve exact medical terminology, to avoid paraphrasing, and Prompt 3 tested a few-shot alternative, providing 2 examples. Prompts 4 and 5 enforced verbatim extraction, explicitly forbidding the correction of misspellings to prevent model over-correction. Prompts 6 and 7 added structural rules enforcing the inclusion of medications/anatomical terms and a single continuous sentence format, respectively. However, stacking these negative and structural constraints within a single instruction frequently induced additional interference, causing the model to sporadically ignore rules or omit key terms.

- **Task Decomposition (Prompts 8-10):** To mitigate constraint overload, we transitioned to an extract-then-generate pipeline. Prompt 8 forced the model to execute an intermediate step: explicitly identifying consecutive two- and three-word medical phrases. Prompt 9 refined this to isolate the primary clinical concern prior to formulating the query. Prompt 10 combined this sequential extraction with few-shot ICL to maximize terminology retention while adhering to brevity limits.

### 3.2. Subtask 2

Subtask 2 evaluates a system's ability to identify clinical note sentences that provide evidence for answering a patient's question. Due to limited development data, we framed this as a prompt-based inference task, instructing models to return the relevant sentence identifiers in a structured JSON format. Each prompt received the clinician question along with the numbered sentences of the clinical note excerpt, and the model was asked to select the identifiers of sentences that provide supporting evidence. We evaluated the following prompt formulations and ensemble strategy:

- **Sentence-Level Classification (Prompts 1-2):** Models evaluated each sentence independently, assigning labels (essential, supplementary, not relevant). This approach introduced instability across sentences and required complex post-hoc aggregation.
- **Case-Level Evidence Selection (Prompts 3-8):** Prompts 3-6 received the entire excerpt and were instructed to output identifiers of relevant sentences directly, emphasizing a "mini-

mal and sufficient" set. However, strict minimality often reduced recall by omitting partial supporting evidence. Prompts 7 and 8 introduced additional reasoning constraints (CoT) to improve transparency, but the increased prompt complexity destabilized output formatting.

- **Recall- and Precision-Oriented Selection (Prompts 9-10):** To explore the trade-off between coverage and specificity, we tested two simplified formulations. Prompt 9 instructed the model to prefer precision over recall, selecting only the most directly relevant sentence identifiers, while Prompt 10 favoured recall over precision, selecting all sentence identifiers that could help answer the clinician's question to reduce the risk of missing relevant evidence.
- **Ensemble Aggregation:** To finalize predictions, we employed a majority voting ensemble across multiple LLMs (Dietterich, 2000). Each model independently generated predictions, and sentence identifiers selected by a minimum of 2/3 models were included in the final output, reducing spurious selections while preserving consistently identified evidence.

### 3.3. Subtask 3

For Subtask 3, the goal was to synthesize a patient-friendly, evidence-grounded answer (maximum 75 words). Each prompt received the patient question, clinician question, and numbered clinical note sentences as input. Because this Subtask heavily evaluates factual consistency and adherence to the source text, our prompt engineering focused on preventing hallucination (Huang et al., 2025) through the following strategies:

- **Few-Shot with Explicit Constraints (Prompts 1-5, 8, 10):** We utilized a 1-shot learning approach, providing an annotated example from the task guidelines. We systematically varied the strictness of length instructions (e.g., "exactly 3 to 5 sentences" vs. "maximum 75 words") and paraphrasing allowances.
- **Multi-Shot Demonstration (Prompt 6):** Expanded ICL by providing two distinct clinical examples, testing whether exposure to diverse specialties improved the model's ability to synthesize answers without exceeding the strict word count.
- **Zero-Shot Abstractive Summarization (Prompt 7):** Removed all in-context examples to test baseline summarization capabilities, relying purely on instructional constraints

to generate faithful, evidence-grounded answers.

- **Implicit Chain-of-Thought / Pre-computation (Prompt 9):** Introduced a cognitive constraint requiring the LLM to silently identify relevant numbered sentences prior to generation. This anchored the model to an extractive mindset without polluting the final output with reasoning traces.
- **High-Density Persona & Negative Constraints (Prompt 11):** Shifted the system persona to a strict “clinical auditor” and densely stacked negative constraints (e.g., “Do NOT infer, generalize, explain, justify, speculate”) to strongly discourage creative text generation.
- **LLM-as-a-Judge:** To better generalize results across model-specific biases, we employed an LLM-as-a-Judge (Zheng et al., 2023) to select the single best candidate answer per case from a pool of outputs with the strongest model-prompt configurations. The judge evaluated candidates according to the following criteria, in priority order: (1) faithfulness: every claim must be supported by the clinical note sentences. (2) medical completeness: covering the key clinical information that answers the question. (3) terminology retention: preserving exact medical terms, procedure names, and diagnoses from the note. (4) conciseness: no filler or hedging, within the 75-word limit.

### 3.4. Subtask 4

Subtask 4 requires systems to link each sentence of a generated answer back to supporting sentences in the clinical note excerpt. Each prompt received the patient question, clinician question, the numbered clinical note sentences, and the generated answer sentences. To ensure accurate sentence-level labels, we preprocessed the texts by explicitly labeling them with structural identifiers prior to prompting. Specifically, each sentence in the source clinical note was prepended with a unique tag (e.g., [1], [2]), and each sentence in the generated answer was similarly tagged. In a variant, answer sentences were additionally prefixed with e.g. [N1], [N2] and [S1], [S2] tags to further disambiguate source and answer identifiers. Some configurations also employed a two-step verification pass, prompting the model to revisit any unaligned answer sentences after the initial output. We framed the model as a clinical evidence alignment expert tasked with generating strictly valid JSON under four core rules: only explicitly supported facts may be cited, references must contain the minimal set that contains all relevant informa-

tion, unsupported sentences must return empty lists, and all answer sentences must be processed.

- **Zero-Shot Alignment (Prompts 1, 10):** Prompt 1 established the baseline with markdown-style rules and explicit-support constraints. Prompt 10 utilized a stricter formulation, explicitly forbidding inference and generalization. However, zero-shot models struggled to consistently differentiate direct evidence from loosely related context.
- **Progressive Few-Shot Scaling (Prompts 2-9):** Prompt 2 introduced a 2-shot setting with clean alignment structures. Prompt 3 expanded to a 4-shot setting incorporating complex cases with high citation density and unsupported answers. Prompt 4 added a fifth conservative example to pull the model toward precise evidence linking. Prompt 5 tested a reordering of the Prompt 4 set to investigate and mitigate recency effects. Prompts 6 and 7 shifted back to 4-shot with alternative cases, while Prompts 8 and 9 extended this to 5-shot configurations, systematically varying example ordering and composition to identify the most effective sequence.
- **Ensemble Aggregation:** Because all few-shot examples were drawn from the 20-case development set, individual prompt configurations risked overfitting. To improve generalization, we searched over candidate model-prompt configurations to construct an ensemble.

## 4. Results

We employed a two-stage evaluation methodology across all four Subtasks. In the initial stage, we conducted extensive validation on the 20-case development set, evaluating a representative, though non-exhaustive, pool of state-of-the-art proprietary models (e.g., Gemini 3 Flash (Doshi and The Gemini Team, 2025), Gemini 2.5 Flash (Comanici et al., 2025), Claude Sonnet 4.5 (Anthropic, 2025), GPT 4.1 (OpenAI, 2025), Grok 4.1 Fast (xAI, 2025)) alongside open-source alternatives (e.g., Llama 3.1 8B Instruct (Llama Team, AI @ Meta, 2024), Qwen 3 (Qwen Team, 2025) and various domain finetunes (Tran et al., 2024)<sup>12</sup>, MedGemma-27B (Sellersgren et al., 2025)). We iteratively adjusted this baseline pool to incorporate newer, higher-performing models as they became available. This validation phase identified our top-performing configurations,

<sup>1</sup><https://huggingface.co/khazarai/Bio-8B-it>

<sup>2</sup><https://huggingface.co/Echelon-AI/Med-Qwen2-7B>

which were subsequently applied to the hidden test set (comprising 47 samples for Subtasks 1-3 and 147 samples for Subtask 4) to generate the official results. To maximize competitive performance, our final submissions featured proprietary models rather than open-source alternatives; we acknowledge this reliance as a limitation of our work. Regarding computational resources, smaller open-source models (<70B parameters) were executed on a SLURM cluster utilizing up to two NVIDIA L4 (24GB) GPUs. All other models were accessed via API endpoints (OpenRouter), incurring a total cost of \$109.45 USD for the duration of the competition. The following subsections detail the validation findings and official performance for each Subtask, with a complete set of validation experiments for Subtasks 1, 2, 3, and 4, including a brief evaluation of very large open-source models, is provided in the Appendix, for Subtask 1, 3, 4.

#### 4.1. Subtask 1

Performance in Subtask 1 is evaluated using an average of four automatic text generation metrics: ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023). Our validation results are presented in Figure 1. Across the Direct Generation phase (Prompts 1-7), performance gradually improved as constraints were refined, though scores remained generally lower than those in the Task Decomposition phase (Prompts 8-10). Task Decomposition consistently achieved the highest average scores across model architectures, suggesting that breaking the task into sequential extraction steps is a highly effective strategy. Based on these validation results, Prompts 8, 9, and 10 were identified as the strongest configurations, with prompt 10 offering the best performance for most proprietary models, indicating the advantage of using few shot examples. In general we note that most proprietary models offer alot more performance and robustness than most open source alternatives.

The official leaderboard results for Subtask 1 are presented in Table 1. Our primary submission utilized Claude Sonnet 4.5 with Prompt 10, combining the sequential extraction pipeline with 2-shot in-context learning (ICL) to maximize terminology retention within the 15-word limit. This approach achieved a score of 19.0, ranking 13th on the leaderboard. A secondary submission using GPT-5.2 with Prompt 8 (extract-then-generate) achieved a score of 16.6, indicating that the in-context examples in Prompt 10 provided a meaningful advantage. Notably, the substantial discrepancy between our internal validation scores and the official test results suggests a potential distribution shift in the test data, an overfitting effect to the highly constrained 20-case development set, or some internal errors

|                   | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   | Average |
|-------------------|------|------|------|------|------|------|------|------|------|------|---------|
| Grok 4.1 Fast     | 74.5 | 88.4 | 81.5 | 79.5 | 90.9 | 90.9 | 91.8 | 93.2 | 93.7 | 94.8 | 87.9    |
| Owen3 Max         | 64.9 | 83.5 | 81.3 | 77.8 | 86.3 | 89.3 | 87.1 | 92.9 | 92.8 | 94.4 | 85.0    |
| GPT 5.2           | 50.2 | 83.4 | 81.0 | 75.0 | 91.7 | 91.5 | 93.3 | 94.7 | 94.2 | 94.6 | 85.0    |
| Gemini 3 Pro      |      |      | 74.5 |      | 82.3 | 82.6 |      | 85.4 | 88.1 | 96.0 | 84.8    |
| Claude Sonnet A.5 | 63.0 | 76.5 | 79.0 | 78.7 | 82.9 | 88.0 | 81.8 | 94.7 | 94.5 | 95.7 | 83.5    |
| Gemini 3 Flash    | 41.9 | 67.3 | 71.1 | 56.1 | 75.0 | 79.8 | 75.8 | 87.7 | 82.7 | 92.4 | 73.0    |
| Claude Sonnet 7B  | 88.0 | 88.2 | 82.5 | 89.9 | 87.6 | 85.5 | 90.1 | 86.8 | 87.4 | 68.6 | 85.5    |
| Med-Owen2 7B      | 49.3 | 73.0 | 69.8 | 61.7 | 75.5 | 76.0 | 80.2 | 88.0 | 86.9 | 80.6 | 74.1    |
| MedGemini 4B      | 68.0 | 46.9 | 58.5 | 71.5 | 76.8 | 65.8 | 76.6 | 86.4 | 79.5 | 89.4 | 71.9    |
| MedGemini 32B     | 61.9 | 66.2 | 75.9 | 69.1 | 85.1 | 71.1 | 68.7 | 60.9 | 77.5 | 70.5 | 70.7    |
| Owen3 1.8B        | 55.8 | 65.9 | 71.6 | 51.0 | 68.9 | 65.4 | 70.5 | 71.9 | 65.5 | 50.9 | 63.7    |
| Llama 3.1 8B      | 61.8 | 73.9 | 75.2 | 71.0 | 82.1 | 80.5 | 81.6 | 85.7 | 85.7 | 84.3 | 78.6    |
| Average           |      |      |      |      |      |      |      |      |      |      |         |

Figure 1: Subtask 1 Validation Results showing the different prompts over several different open and close source models. The scores represent the official evaluation metrics of the average of ROUGE, BERTScore, AlignScore, and MEDCON.

on our part.

| Model / Team           | Rank     | Score       |
|------------------------|----------|-------------|
| <i>Our Submissions</i> |          |             |
| Sonnet-4.5 Prompt 10   | 13       | 19.0        |
| GPT-5.2 Prompt 8       | —        | 16.6        |
| <i>Leaderboard</i>     |          |             |
| Best Competitor        | <b>1</b> | <b>31.2</b> |
| Median                 | 7        | 25.6        |

Table 1: Subtask 1 Official Results. The score represents the average of ROUGE, BERTScore, AlignScore, and MEDCON.

#### 4.2. Subtask 2

Performance in Subtask 2 is evaluated using the Strict Micro F1 score to measure a system’s ability to accurately identify the minimal and sufficient set of relevant evidence sentences. Validation results (Figure 2) highlight a distinct performance gap between proprietary and open-source models. Proprietary models consistently achieved the highest scores, typically exceeding 50.0 Micro F1 and demonstrating relative stability across various prompt configurations. In contrast, open-source models produced substantially lower scores and exhibited higher sensitivity to prompt variations. Further we note many cases of complete failure (F1 score of zero), where the model/prompt configuration would generate responses which were unparseable. In terms of our prompting approaches, we noticed no clear winners in terms of prompts, with various models having various

optimal prompts, with the exception of prompt 10, obtaining top performance for most propriety models.

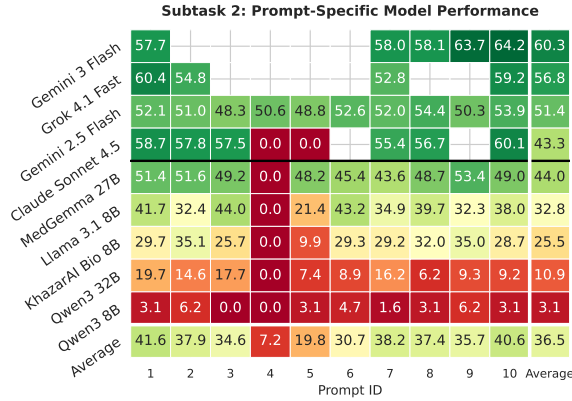


Figure 2: Subtask 2 Validation Results over the Strict Micro F1 metric

Consequently, our final system relied on an ensemble of the strongest proprietary models. The primary submission, a majority voting ensemble comprising Gemini 3 Flash Preview, Grok 4.1 Fast, and Claude Sonnet 4.5 (all using prompt 10), achieved a Strict Micro F1 score of 58.8, ranking 11th on the leaderboard (Table 2). This performance is near the median system score (59.8) and remains competitive with the top-performing system (63.7). An alternative single-model submission (Grok 4 Fast Prompt 10) achieved a score of 56.0. The minimal variance between the ensemble and single-model runs demonstrates the stability of the prompting strategy, suggesting that a single model is highly reliable and rendering more computationally expensive ensembling strategies unnecessary.

| Model / Team            | Rank | Strict Micro F1 |
|-------------------------|------|-----------------|
| <i>Our Submissions</i>  |      |                 |
| Ensemble Top 3          | 11   | 58.8            |
| Grok 4 Fast (Prompt 10) | –    | 56.0            |
| <i>Leaderboard</i>      |      |                 |
| Best Competitor         | 1    | 63.7            |
| Median                  | 7/8  | 59.8            |

Table 2: Subtask 2 Performance: Strict Micro F1 on ArchEHR-QA.

### 4.3. Subtask 3

Performance in Subtask 3 is evaluated using an average score over the metrics: BLEU, ROUGE, SARI (Xu et al., 2016), BERTScore, AlignScore, and MEDCON. Validation results (Figure 3) revealed that performance narrowly varied across

all eleven prompts. This indicates that for patient-friendly answer generation, models are relatively robust to prompt variation, making model selection the primary driver of performance. Gemini 2.5 Flash and Claude Sonnet 4.5 consistently outperformed other models, while open-source models like MedGemma-27B and a domain fine-tuned Qwen 3-8B<sup>3</sup> performed competitively, narrowing the performance gap observed in other Subtasks.

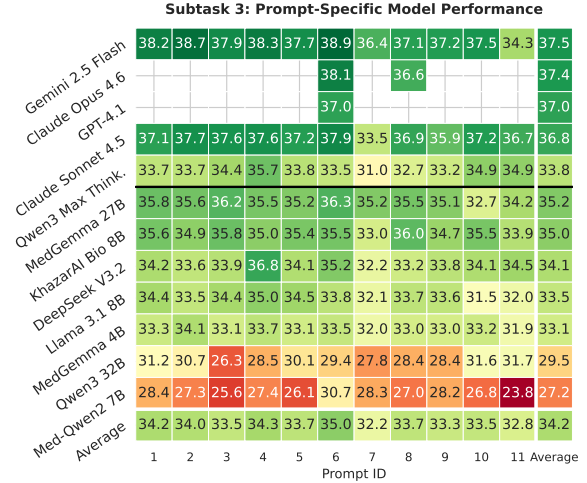


Figure 3: Subtask 3 Validation Results using an average score of the metrics: BLEU, ROUGE, SARI, BERTScore, AlignScore, and MEDCON

Given the marginal prompt-based variance, our submission strategy prioritized model diversity and high-quality demonstrations. Our first submission utilized an LLM-as-a-Judge (GPT-4.1) over the strongest model-prompt configurations: Gemini 2.5 Flash, Claude Sonnet 4.5, and Claude Opus 4.6 on prompt 6. This approach ranked 3rd on the official leaderboard with a score of 35.6 (Table 3), closely trailing the top competitor (36.3). Single-model submissions also demonstrated strong performance: Gemini 2.5 Flash (Prompt 6, multi-shot) achieved 35.5, while Claude Sonnet 4.5 (Prompt 2, 1-shot explicit constraints) scored 33.9. This suggests that for this generation task, the LLM-as-a-Judge does not add much value relative to the overhead. The narrow 0.7-point gap between our best ensemble and the winning submission underscores the viability of prompt-based approaches for simplified medical text generation.

### 4.4. Subtask 4

Performance in Subtask 4 is evaluated using sentence-level precision, recall, and Micro F1 metrics over predicted alignment links. Validation results (Figure 4) indicated that using few shot prompt-

<sup>3</sup><https://huggingface.co/khazarai/Bio-8B-it>



| Model / Team                | Rank | Score |
|-----------------------------|------|-------|
| <i>Our Submissions</i>      |      |       |
| LLM-as-a-Judge 4 Models     | 3    | 35.6  |
| Gemini-2.5-Flash (Prompt 6) | –    | 35.5  |
| Sonnet-4.5 (Prompt 2)       | –    | 33.9  |
| <i>Leaderboard</i>          |      |       |
| Best Competitor             | 1    | 36.3  |
| Leaderboard Median          | 7    | 32.9  |

Table 3: Subtask 3 Performance Comparison. The Overall metric is a composite score evaluating simplified medical text generation, incorporating BLEU, ROUGE-Lsum, SARI, BERTScore, Align-Score, and MEDCON (UMLS).

ing (Prompts 2-9), often gave the open source models a clear benefit, with most models obtaining their best scores between prompts 5-7. However, because all few-shot examples were drawn from the evaluation development set, we believed that this might reflect partial overfitting. In general, proprietary models significantly outperformed open-source alternatives. Google’s MedGemma-27B was a notable exception (average 81.0), though it still trailed the top proprietary models.

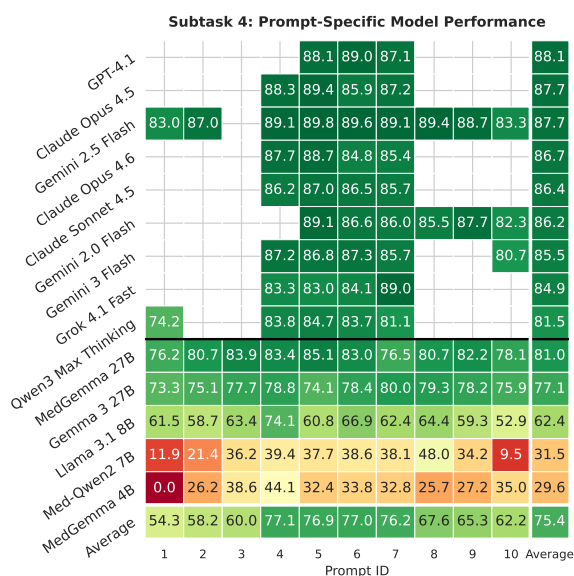


Figure 4: Subtask 4 Validation Results

To maximize robustness, we utilized an ensemble search over the strongest model-prompt combinations (Prompts 5-7), with the final submissions using Prompts 5 and 6. The final ensemble composition was selected based solely on Micro F1 performance on the 20-case development set, with no subsequent adjustment made prior to or following test set submission. Our primary submissions, a 3-model majority voting ensemble (Gemini 2.5 Flash, Claude Opus 4.6, Gemini 2.0 Flash) with *Prompt 5*,

achieved 1st place on the leaderboard with a Micro F1 of 81.5 (Table 4). Our second submission used a 4-model ensemble combining Gemini 2.5 Flash and GPT-4.1, each evaluated with *Prompts 5 and 6*, resulting in four model-prompt configurations and achieving 81.3. Finally, a strictly constrained zero-shot single-model configuration (Gemini 2.5 Flash, Prompt 10) achieved 81.2, demonstrating that carefully engineered zero-shot prompting can perform competitively with few-shot ensembles. Although the performance differences between our approaches is minimal, the margin between the 1st and 4th place submission was 1.2 Micro F1, indicating that these marginal changes affect the ranking of the systems. Furthermore, despite earlier concerns about potential overfitting due to few-shot examples drawn from the development set, the similarity between validation and official results suggests the overfitting was minimal.

| Model / Team                 | Rank | Micro F1 |
|------------------------------|------|----------|
| <i>Our Submissions</i>       |      |          |
| Ensemble 3 models            | 1    | 81.5     |
| Ensemble 4 models            | –    | 81.3     |
| Gemini 2.5 Flash (Prompt 10) | –    | 81.2     |
| <i>Leaderboard</i>           |      |          |
| Best Competitor              | 2    | 81.3     |
| Median                       | 8/9  | 77.5     |

Table 4: Subtask 4 Performance: Micro F1 on ArchEHR-QA.

## 5. Error Analysis

**Subtask 1: Question Interpretation Error Analysis.** The model occasionally failed to transform patient narratives into professional clinical queries. In Case 129, it generated an informal, query with lots of spelling errors (“*Why is he not eatin feeling weak and shakey loosin weight...*”) rather than the required concise clinical formulation (“*What explains his weakness, weight loss, and poor appetite...*”). Upon reflection this is a model directly following our instructions where we ask specifically to not rephrase or correct spelling errors.

**Subtask 2: Evidence Retrieval Error Analysis.** Retrieval primarily suffered from over-selection and missed evidence. In over-selection cases (e.g., Case 19), the system successfully retrieved the gold annotations but included numerous false positives containing irrelevant procedural background due to an over-reliance on broad semantic similarity. Notably, our attempts to mitigate this using stricter filtering and ensemble methods have not fully resolved the issue. Conversely, in missed evidence



cases (e.g., Case 20), the model hyper-focused on explicit diagnostic terminology, completely overlooking relevant sentences describing treatments. Addressing this requires non-trivial improvements in semantic reasoning.

**Subtask 3: Answer Generation Error Analysis.** Generation errors typically involved medical hallucinations or context misinterpretation. In Case 21, the model correctly identified a patient's cirrhosis but hallucinated plausible yet entirely unsupported complications (*portal gastropathy*, *Budd-Chiari syndrome*). In Case 29, it ignored the primary clinical note regarding deep vein thrombosis treatment, instead generating an unrelated discussion about Coumadin and alcohol use. These issues emphasize the need for stronger grounding mechanisms to tightly couple generation with the retrieved evidence.

**Subtask 4: Evidence Alignment Error Analysis.** The model demonstrated difficulty with multi-evidence reasoning, where multiple context sentences jointly support a single claim. In Case 3, the model correctly cited one supporting sentence but missed a second complementary citation detailing warning symptoms, largely because the missing sentence utilized different terminology. Mitigating these incomplete citations will require prompting and retrieval strategies that explicitly encourage aggregating complementary evidence.

## 6. Discussion

In this work, we aimed to evaluate the limits of LLMs in the extreme low-resource biomedical setting of the ArchEHR competition, relying primarily on prompt engineering and model selection rather than supervised fine-tuning. Our findings highlight several key dynamics regarding model scaling, open-source viability, and the practical trade-offs of ensemble methodologies in clinical applications.

Addressing the viability of open-weights models for secure clinical deployment (RQ1), internal validation highlighted the impressive capabilities of open-source, domain-adapted models. Notably, MedGemma 3 27B performed exceptionally well, closely trailing large proprietary counterparts despite its relatively small parameter count. This shrinking performance gap is critical for the healthcare domain, where data privacy regulations often strictly prohibit sending sensitive patient data to external API endpoints. Locally deployable domain tuned open-source models are often the only option in these settings; however, our findings confirm that more work is needed to extract optimal performance from these smaller models.

Regarding the effectiveness of prompt engineering (RQ2), our validation phase revealed that the impact of specific techniques is inversely proportional

to baseline model capability. State-of-the-art proprietary models generally demonstrated high robustness, yielding less performance variance across different prompt templates while maintaining higher relative performance, with several prompts often achieving similarly top-tier results. Conversely, open-source LLMs exhibited significant, model-dependent performance swings (e.g., the high variance between Qwen 3-32B and MedGemma architectures in Subtask 1, though Subtask 3 proved an exception). However, even for the most robust models, careful prompt tuning remained a necessity for enforcing strict structural constraints, such as rectifying the consistent JSON formatting failures observed in Subtask 2. Furthermore, it is important to note that poorly constructed prompts will still degrade the performance of proprietary models, as evidenced by Prompt 1 in Subtask 1. In general, regarding proprietary models, once a baseline of high performance is reached, further prompt engineering yields diminishing returns. This is aligned with the existing literature (Zhuo et al., 2024).

To maximize our competitive standing, we utilized ensembling techniques across the pipeline, including majority voting and LLM-as-a-judge frameworks. While these strategies consistently improved robustness and yielded positive results, the absolute performance gains were often marginal. In a shared task competition, fractional improvements are vital for leaderboard positioning. However, from a real-world clinical deployment perspective, these techniques introduce severe operational bottlenecks. Running multiple inferences per query effectively multiplies operational costs or latency by a factor of the ensemble size. For live, scalable health systems, the marginal gains of ensembling are not worth the resulting computational inefficiency.

Broadly, the ArchEHR competition demonstrates that LLMs are highly competitive in this low-resource clinical setting, though their efficacy varies significantly by task type. For Subtask 3 (answer generation), LLMs remain the natural and dominant choice, excelling at abstractive synthesis. Surprisingly, our prompt-based ensemble also achieved state-of-the-art results in Subtask 4 (evidence citation), which we initially hypothesized would be dominated by fine-tuned systems. Conversely, the leaderboard standings for Subtask 2 suggest that other approaches may be more optimal over prompt-based LLMs for pure evidence-span retrieval, if the best competitor uses other techniques. Finally, we note that our official Subtask 1 performance was an anomaly compared to our internal validation; we believe there remains substantial, untapped potential for LLM-driven query transformation in this space.

## 7. Conclusion

This study investigated the efficacy of LLMs in the extreme low-resource clinical setting of the ArchEHR competition. By systematically evaluating zero-shot, few-shot, and CoT prompting strategies alongside ensemble methodologies, we demonstrated that state-of-the-art LLMs can achieve highly competitive performance across complex biomedical NLP tasks without any supervised fine-tuning. Notably, our prompt-based ensemble approach achieved first place in Subtask 4 (evidence citation alignment) and third place in Subtask 3 (patient-friendly answer generation), proving that LLMs are naturally suited for both abstractive synthesis and strict evidence-linking. Conversely, our lower performance in Subtasks 1 and 2 indicates that pure evidence-span retrieval and strict query formulation remain challenging for generative, prompt-only architectures.

Our findings also highlight critical dynamics for the real-world deployment of clinical AI. While massive proprietary models exhibited strong robustness to prompt variations, open-source models, particularly MedGemma 3 27B, demonstrated highly competitive capabilities with certain prompts. This narrowing performance gap is a crucial development for healthcare environments bound by strict data privacy frameworks like GDPR, where local deployment is mandatory. Finally, while ensembling techniques such as majority voting and LLM-as-a-judge provided the marginal gains necessary for competitive leaderboard positioning, their associated computational costs and latency make them inefficient for scalable, real-world clinical deployment.

## 8. Limitations

While our work showed good results, this study has several notable limitations. First, due to the extreme low-resource nature of the shared task (only 20 development cases), there is a persistent risk that our few-shot prompts and ensemble configurations are partially overfit to the development distribution, despite our efforts to utilize varying case structures. This impacts the conclusions drawn from the validation results. Additionally, our uniform use of temperature 0.0 and top-p of 0.95 across all models, while intended to maximize deterministic behavior, may not be optimal for all architectures evaluated, particularly reasoning-oriented models that are designed to benefit from some degree of stochasticity.

Second, our evaluation of proprietary LLMs was not exhaustive, and was not run on all prompt configurations. The financial constraints associated with commercial API costs limited our ability to perform evaluate every prompt variation across all

available proprietary models. Consequently, the optimal prompt-model combinations reported may reflect a local maximum rather than absolute peak performance.

Furthermore, our system heavily relies on proprietary, closed-weights models (e.g., Gemini and Claude) accessed via API endpoints to achieve top-tier performance. This introduces reproducibility challenges, as the underlying model weights may be silently updated by the providers. And in GDPR strict scenarios this cannot be permitted. Finally, the evaluation was restricted to English clinical notes; the efficacy of these prompting strategies on multilingual or non-English EHRs remains unverified.

## 9. Ethical Considerations

The deployment of generative LLMs in clinical settings carries profound ethical implications, primarily concerning patient safety and data privacy. Generative models are inherently prone to hallucination. While our system achieved high precision in evidence alignment (Subtask 4) and grounding (Subtask 3), the risk of clinical hallucination cannot be entirely eliminated through prompt engineering alone. Therefore, systems like ours must strictly be deployed as “human-in-the-loop” assistive tools to reduce clinician cognitive load, rather than as autonomous diagnostic or patient-facing decision-makers.

Additionally, our highest-performing submissions rely on sending clinical text to external APIs. In real-world healthcare environments governed by privacy frameworks such as HIPAA and GDPR, transmitting sensitive Protected Health Information to external servers is often legally prohibited. While we demonstrated that local, open-weights models like MedGemma 3 27B show strong promise, bridging the performance gap between secure local models and proprietary systems remains a critical ethical and technical imperative for the field.

## 10. Acknowledgments

This work was funded by FEDER - Fundo Europeu de Desenvolvimento Regional funds through Programa Regional do Centro, within project CENTRO2030-FEDER-02595400 and by the Foundation for Science and Technology (FCT) through the contract <https://doi.org/10.54499/UID/00127/2025>.

Richard A. A. Jonker is funded by the FCT doctoral grant PRT/BD/154792/2023, with DOI identifier <https://doi.org/10.54499/PRT/BD/154792/2023>. Alexander Christiansen and Alexandros Maniatis are funded by the Danish

Data Science Academy (DDSA) under grant numbers 2026-6484 and 2026-6480, respectively.

## 11. Bibliographical References

- 104th United States Congress. 1996. [Health Insurance Portability and Accountability Act of 1996 \(HIPAA\)](#). Public Law 104-191, 110 Stat. 1936.
- Anthropic. 2025. [Introducing claude sonnet 4.5](#). Anthropic News. Accessed: 2026-03-10.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. Unleashing the potential of prompt engineering for large language models. *Patterns*, 6(6).
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.
- Tulsee Doshi and The Gemini Team. 2025. [Gemini 3 flash: frontier intelligence built for speed](#). Google The Keyword Blog. Accessed: 2026-03-10.
- European Union. 2016. [Regulation \(EU\) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC \(General Data Protection Regulation\)](#). Official Journal of the European Union, L 119, 1-88.
- Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- 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.
- Llama Team, AI @ Meta. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2023. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*, pages 24457–24477. PMLR.
- OpenAI. 2025. [Introducing gpt-4.1 in the api](#). OpenAI Blog. Accessed: 2026-03-10.
- Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 1.
- Maximilian Schall and Gerard de Melo. 2025. [The hidden cost of structure: How constrained decoding affects language model performance](#). In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1074–1084, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.
- 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.
- Sarvesh Soni and Dina Demner-Fushman. 2026a. [A dataset for addressing patient's information needs related to clinical course of hospitalization](#). *Scientific Data*.

- 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.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.
- Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. [Bioinstruct: instruction tuning of large language models for biomedical natural language processing](#). *Journal of the American Medical Informatics Association*.
- 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 (NeurIPS)*.
- xAI. 2025. [Grok 4.1 fast and agent tools api](#). xAI News. Accessed: 2026-03-10.
- 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.
- Wen-wai Yim, Yajuan 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. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. [Least-to-most prompting enables complex reasoning in large language models](#). In *The Eleventh International Conference on Learning Representations*.
- Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. [ProSA: Assessing and understanding the prompt sensitivity of LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

## A. Additional Experiments

In this appendix, we present the full set of experiments conducted on the development set for Subtasks 1, 3, and 4 (Figures 5, 6, and 7, respectively). Notably, these extended validation results demonstrate that very large open-source models (such as Qwen 3 235B, GLM4.7, and Deepseek V3.2) performed exceptionally well, achieving results highly competitive with proprietary models. However, we chose not to emphasize these models in the main text due to their prohibitive local hardware requirements. Running these models locally requires substantial compute resources (often exceeding 140GB of VRAM), which is impractical for most standard biomedical deployments. Because of this limitation, we accessed these open-source models via API endpoints during testing; however, for API-based applications, we opted to focus our primary analysis on established proprietary models. For these practical reasons, we did not investigate these massive open-source models further in the main study.



| Subtask 1: Prompt-Specific Model Performance (All Models) |           |      |      |      |      |      |      |      |      |      |         |
|-----------------------------------------------------------|-----------|------|------|------|------|------|------|------|------|------|---------|
| Grok 4.1 Fast                                             | 74.5      | 88.4 | 81.5 | 79.5 | 90.9 | 90.9 | 91.8 | 93.2 | 93.7 | 94.8 | 87.9    |
| Qwen3 Max                                                 | 64.9      | 83.5 | 81.3 | 77.8 | 86.3 | 89.3 | 87.1 | 92.9 | 92.8 | 94.4 | 85.0    |
| GPT-5.2                                                   | 50.2      | 83.4 | 81.0 | 75.0 | 91.7 | 91.5 | 93.3 | 94.7 | 94.2 | 94.6 | 85.0    |
| Gemini 3 Pro                                              |           |      | 74.5 |      | 82.3 | 82.6 |      | 85.4 | 88.1 | 96.0 | 84.8    |
| Claude Sonnet 4.5                                         | 63.0      | 76.5 | 79.0 | 78.7 | 82.9 | 88.0 | 81.8 | 94.7 | 94.5 | 95.7 | 83.5    |
| Claude Opus 4.5                                           | 64.2      | 80.1 | 73.7 | 79.4 | 84.5 | 84.7 | 83.8 | 86.2 | 88.7 | 92.7 | 81.8    |
| Kimi K2.5                                                 | 55.4      | 78.6 | 81.3 | 71.0 | 74.5 | 81.6 |      | 78.0 | 79.1 | 59.2 | 73.2    |
| Gemini 3 Flash                                            | 41.9      | 67.3 | 71.1 | 56.1 | 75.0 | 79.8 | 75.8 | 87.7 | 82.7 | 92.4 | 73.0    |
| Med-Qwen2 7B                                              | 88.0      | 88.2 | 82.5 | 89.9 | 87.6 | 85.5 | 90.1 | 86.8 | 87.4 | 68.6 | 85.5    |
| Qwen3 235B                                                | 60.2      | 82.0 | 77.7 | 72.0 | 88.3 | 86.9 | 89.8 | 92.6 | 92.6 | 93.8 | 83.6    |
| GLM-4.7                                                   | 59.5      | 80.7 | 75.7 | 71.5 | 87.3 | 87.4 | 85.5 | 87.0 | 91.2 | 94.2 | 82.0    |
| DeepSeek V3.2                                             | 56.8      | 76.4 | 72.4 | 68.0 | 84.4 | 83.9 | 84.5 | 89.8 | 83.4 | 91.8 | 79.1    |
| CURE-MED 14B                                              | 72.9      | 76.7 | 69.8 | 72.9 | 82.9 | 79.5 | 79.5 | 83.4 | 81.0 | 85.2 | 78.4    |
| GPT OSS 120B                                              | 52.1      | 59.5 | 67.5 | 72.0 | 82.9 | 73.0 | 84.0 | 90.7 | 87.7 | 92.5 | 76.2    |
| MedGemma 27B                                              | 49.3      | 73.0 | 69.8 | 61.7 | 75.5 | 76.0 | 80.2 | 88.0 | 86.9 | 80.6 | 74.1    |
| MedGemma 1.5 4B                                           | 69.5      | 75.6 | 74.7 | 72.2 | 74.6 | 81.7 | 78.5 | 80.9 | 62.4 | 69.4 | 74.0    |
| DeepSeek R1                                               | 50.8      | 69.4 | 60.9 |      | 81.3 | 80.5 |      | 84.0 |      | 81.2 | 72.6    |
| Llama 4 Maverick                                          | 61.3      | 66.7 | 70.9 | 62.3 | 75.5 | 71.9 | 75.5 | 80.6 | 79.0 | 81.3 | 72.5    |
| MedGemma 27B IT                                           | 58.5      | 63.5 | 68.2 | 53.3 | 74.0 | 70.7 | 72.4 | 90.5 | 86.5 | 84.5 | 72.2    |
| Lingshu 7B                                                | 70.0      | 70.2 | 69.0 | 57.0 | 76.5 | 77.3 | 72.2 | 73.4 | 72.5 | 82.2 | 72.0    |
| MedGemma 4B                                               | 68.0      | 46.9 | 58.5 | 71.5 | 76.8 | 65.8 | 76.6 | 86.4 | 79.5 | 89.4 | 71.9    |
| Lingshu 32B                                               | 62.9      | 68.5 | 65.0 | 67.1 | 76.4 | 71.6 | 76.5 | 79.3 | 75.6 | 75.6 | 71.9    |
| Qwen3 32B                                                 | 61.9      | 66.2 | 75.9 | 69.1 | 85.1 | 71.1 | 68.7 | 60.9 | 77.5 | 70.5 | 70.7    |
| MedGemma Heretic 27B                                      | 48.0      | 69.1 | 66.3 | 53.7 | 77.5 | 76.1 | 79.6 | 82.7 | 84.5 | 67.7 | 70.5    |
| Llama 3.3 70B                                             | 62.8      | 67.4 | 67.9 | 67.9 | 73.9 | 70.0 | 75.5 | 68.2 | 70.2 | 76.8 | 70.0    |
| Fine-tuned Qwen1 7B                                       | 66.6      | 76.7 | 71.8 | 66.5 | 69.2 | 83.5 | 66.8 | 75.3 | 78.5 | 32.5 | 68.7    |
| Gemma 3 27B                                               | 47.4      | 60.6 | 58.8 | 48.4 | 72.4 | 65.8 | 74.2 | 82.9 | 75.9 | 61.6 | 64.8    |
| Llama 3.1 8B                                              | 55.8      | 65.9 | 71.6 | 51.0 | 68.9 | 65.4 | 70.5 | 71.9 | 65.5 | 50.9 | 63.7    |
| Qwen3 1.7B Med                                            | 51.8      | 57.2 | 51.9 | 54.3 | 69.3 | 49.1 | 64.0 | 57.2 | 52.1 | 50.0 | 55.7    |
| Qwen3 0.6B Med                                            | 54.5      | 53.6 | 71.1 | 54.6 | 53.4 | 56.6 | 61.3 | 47.3 | 36.5 | 30.3 | 51.9    |
| BioMistral 7B                                             | 60.0      | 55.2 | 50.9 | 46.1 | 55.4 | 58.0 | 64.1 | 49.2 | 36.5 | 9.8  | 48.5    |
| II Medical 8B                                             | 43.0      | 49.2 | 64.1 | 30.8 | 63.6 | 52.0 | 54.5 | 38.8 | 27.7 | 56.3 | 48.0    |
| Nemotron 3 Nano 30B                                       | 51.6      | 64.9 | 53.2 | 50.5 | 45.6 | 38.7 | 11.2 | 19.2 | 23.3 | 16.5 | 37.5    |
| Ministral 14B                                             | 32.2      | 22.8 | 69.3 | 35.3 | 33.0 | 30.8 | 20.5 | 18.6 | 13.7 | 39.2 | 31.6    |
| Average                                                   | 58.5      | 68.6 | 70.0 | 63.7 | 75.3 | 73.4 | 73.2 | 75.8 | 73.4 | 72.1 | 70.6    |
|                                                           | 1         | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   | Average |
|                                                           | Prompt ID |      |      |      |      |      |      |      |      |      |         |

Figure 5: Subtask 1 Validation Results (All Models)



| Subtask 3: Prompt-Specific Model Performance (All Models) |           |      |      |      |      |      |      |      |      |      |      |         |
|-----------------------------------------------------------|-----------|------|------|------|------|------|------|------|------|------|------|---------|
| Gemini 2.5 Flash                                          | 38.2      | 38.7 | 37.9 | 38.3 | 37.7 | 38.9 | 36.4 | 37.1 | 37.2 | 37.5 | 34.3 | 37.5    |
| Claude Opus 4.6                                           |           |      |      |      |      | 38.1 |      | 36.6 |      |      |      | 37.4    |
| GPT-4.1                                                   |           |      |      |      |      | 37.0 |      |      |      |      |      | 37.0    |
| Claude Sonnet 4.5                                         | 37.1      | 37.7 | 37.6 | 37.6 | 37.2 | 37.9 | 33.5 | 36.9 | 35.9 | 37.2 | 36.7 | 36.8    |
| Gemini 3 Flash                                            |           |      |      |      |      | 36.6 |      |      |      |      |      | 36.6    |
| Grok 4.1 Fast                                             |           | 35.0 |      |      |      | 34.2 |      |      |      |      |      | 34.6    |
| Qwen3 Max Thinking                                        | 33.7      | 33.7 | 34.4 | 35.7 | 33.8 | 33.5 | 31.0 | 32.7 | 33.2 | 34.9 | 34.9 | 33.8    |
| GPT-5 Mini                                                |           |      |      |      |      | 32.8 |      |      |      |      |      | 32.8    |
| GPT-5.2                                                   |           |      |      |      |      | 32.1 |      | 32.1 | 34.0 |      |      | 32.7    |
| GLM-4.6V                                                  |           | 36.8 |      |      |      |      |      |      |      |      |      | 36.8    |
| MedGemma 27B                                              | 35.8      | 35.6 | 36.2 | 35.5 | 35.2 | 36.3 | 35.2 | 35.5 | 35.1 | 32.7 | 34.2 | 35.2    |
| KhazarAI Bio 8B                                           | 35.6      | 34.9 | 35.8 | 35.0 | 35.4 | 35.5 | 33.0 | 36.0 | 34.7 | 35.5 | 33.9 | 35.0    |
| DeepSeek V3.2                                             | 34.2      | 33.6 | 33.9 | 36.8 | 34.1 | 35.2 | 32.2 | 33.2 | 33.8 | 34.1 | 34.5 | 34.1    |
| Llama 3.1 8B                                              | 34.4      | 33.5 | 34.4 | 35.0 | 34.5 | 33.8 | 32.1 | 33.7 | 33.6 | 31.5 | 32.0 | 33.5    |
| Gemma 3 27B                                               | 34.7      | 34.4 | 31.1 | 33.3 | 32.1 | 33.9 | 31.7 | 34.0 | 32.7 | 34.6 | 35.0 | 33.4    |
| MedGemma 4B                                               | 33.3      | 34.1 | 33.1 | 33.7 | 33.1 | 33.5 | 32.0 | 33.0 | 33.0 | 33.2 | 31.9 | 33.1    |
| MedGemma 1.5 4B                                           | 32.1      | 31.5 | 30.9 | 30.4 | 30.5 | 31.5 | 33.3 | 30.9 | 30.8 | 32.3 | 29.9 | 31.3    |
| Qwen3 8B                                                  | 31.9      | 31.7 | 30.9 | 31.8 | 30.4 | 30.9 | 29.4 | 31.0 | 30.2 | 31.4 | 32.1 | 31.1    |
| Qwen3 32B                                                 | 31.2      | 30.7 | 26.3 | 28.5 | 30.1 | 29.4 | 27.8 | 28.4 | 28.4 | 31.6 | 31.7 | 29.5    |
| BioMistral 7B                                             | 28.9      | 29.1 | 29.2 | 29.0 | 29.1 | 30.6 | 27.5 | 28.9 | 27.7 | 28.4 | 27.9 | 28.8    |
| Med-Qwen2 7B                                              | 28.4      | 27.3 | 25.6 | 27.4 | 26.1 | 30.7 | 28.3 | 27.0 | 28.2 | 26.8 | 23.8 | 27.2    |
| Average                                                   | 33.5      | 33.6 | 32.7 | 33.4 | 32.8 | 34.1 | 31.7 | 32.9 | 32.6 | 33.0 | 32.3 | 33.7    |
|                                                           | 1         | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   | 11   | Average |
|                                                           | Prompt ID |      |      |      |      |      |      |      |      |      |      |         |

Figure 6: Subtask 3 Validation Results (All Models)

Subtask 4: Prompt-Specific Model Performance (All Models)

|                       |           |      |      |      |      |      |      |      |      |      |  |         |
|-----------------------|-----------|------|------|------|------|------|------|------|------|------|--|---------|
| GPT-4.1               |           |      |      |      | 88.1 | 89.0 | 87.1 |      |      |      |  | 88.1    |
| Claude Opus 4.5       |           |      |      | 88.3 | 89.4 | 85.9 | 87.2 |      |      |      |  | 87.7    |
| Gemini 2.5 Flash      | 83.0      | 87.0 |      | 89.1 | 89.8 | 89.6 | 89.1 | 89.4 | 88.7 | 83.3 |  | 87.7    |
| Claude Opus 4.6       |           |      |      | 87.7 | 88.7 | 84.8 | 85.4 |      |      |      |  | 86.7    |
| Claude Sonnet 4.5     |           |      |      | 86.2 | 87.0 | 86.5 | 85.7 |      |      |      |  | 86.4    |
| Gemini 2.0 Flash      |           |      |      |      | 89.1 | 86.6 | 86.0 | 85.5 | 87.7 | 82.3 |  | 86.2    |
| Qwen3.5 Plus          |           |      |      |      | 86.1 |      |      |      |      |      |  | 86.1    |
| Grok 4 Fast           |           |      |      | 87.3 | 87.4 | 85.2 | 83.6 |      |      |      |  | 85.9    |
| Claude Sonnet 4       |           |      |      | 86.7 | 86.2 | 84.5 | 85.9 |      |      |      |  | 85.8    |
| Claude Sonnet 4.6     |           |      |      | 87.3 | 85.1 | 83.5 | 86.4 |      |      |      |  | 85.6    |
| Gemini 3 Flash        |           |      |      | 87.2 | 86.8 | 87.3 | 85.7 |      |      | 80.7 |  | 85.5    |
| Gemini 3.1 Pro        |           |      |      |      | 85.0 |      |      |      |      |      |  | 85.0    |
| Grok 4.1 Fast         |           |      |      | 83.3 | 83.0 | 84.1 | 89.0 |      |      |      |  | 84.9    |
| GPT-5                 |           |      |      |      | 83.5 | 83.0 | 87.9 |      |      |      |  | 84.8    |
| Gemini 2.5 Flash Lite |           |      |      | 85.5 | 83.0 |      |      |      |      |      |  | 84.2    |
| MiniMax M2.1          |           |      |      |      | 83.4 |      |      |      |      |      |  | 83.4    |
| GPT-4.1 Mini          |           |      |      |      | 83.2 |      |      |      |      |      |  | 83.2    |
| GPT-5.1               |           |      |      |      | 83.0 |      |      |      |      |      |  | 83.0    |
| MiniMax M2.5          |           |      |      |      | 82.0 |      |      |      |      |      |  | 82.0    |
| Grok 4                |           |      |      | 79.1 | 84.9 |      |      |      |      |      |  | 82.0    |
| Claude Haiku 4.5      |           |      |      |      | 84.0 | 80.5 | 80.4 |      |      |      |  | 81.6    |
| Qwen3 Max Thinking    | 74.2      |      |      | 83.8 | 84.7 | 83.7 | 81.1 |      |      |      |  | 81.5    |
| GPT-5 Mini            |           |      |      |      | 79.5 | 78.4 | 80.6 |      |      |      |  | 79.5    |
| GPT-5.2               |           |      |      |      | 74.9 |      |      |      |      |      |  | 74.9    |
| Gemini 2.5 Pro        |           |      |      |      | 74.9 |      |      |      |      |      |  | 74.9    |
| GPT-5 Nano            |           |      |      |      | 68.7 |      |      |      |      |      |  | 68.7    |
| Gemini 3 Pro          |           |      |      |      | 68.1 |      |      |      |      |      |  | 68.1    |
| Kimi K2.5             |           |      |      | 67.6 |      |      |      |      |      |      |  | 67.6    |
| Kimi K2 Thinking      |           |      |      | 51.8 |      |      |      |      |      |      |  | 51.8    |
| Qwen3.5 397B          |           |      |      |      | 86.7 |      |      |      |      |      |  | 86.7    |
| DeepSeek V3           |           |      |      |      | 85.9 |      |      |      |      |      |  | 85.9    |
| DeepSeek V3.2         |           |      |      |      | 84.0 |      |      |      |      |      |  | 84.0    |
| MiMo V2 Flash         |           |      |      |      | 83.4 |      |      |      |      |      |  | 83.4    |
| Qwen3 235B            |           |      |      |      | 83.3 |      |      |      |      |      |  | 83.3    |
| MedGemma 27B          | 76.2      | 80.7 | 83.9 | 83.4 | 85.1 | 83.0 | 76.5 | 80.7 | 82.2 | 78.1 |  | 81.0    |
| Gemma 3 27B           | 73.3      | 75.1 | 77.7 | 78.8 | 74.1 | 78.4 | 80.0 | 79.3 | 78.2 | 75.9 |  | 77.1    |
| Nemotron 3 Nano 30B   |           |      |      |      | 78.9 | 75.6 | 73.7 |      |      |      |  | 76.1    |
| Llama 3.3 70B         |           |      |      |      | 74.9 |      |      |      |      |      |  | 74.9    |
| Llama 4 Scout         |           |      |      |      | 66.7 |      |      |      |      |      |  | 66.7    |
| GLM-4.7 Flash         |           |      |      |      | 63.3 |      |      |      |      |      |  | 63.3    |
| Llama 3.1 8B          | 61.5      | 58.7 | 63.4 | 74.1 | 60.8 | 66.9 | 62.4 | 64.4 | 59.3 | 52.9 |  | 62.4    |
| MedGemma 1.5 4B       | 38.9      | 56.4 | 50.5 | 59.2 | 49.8 | 61.1 | 56.0 | 61.5 | 51.3 | 41.9 |  | 52.7    |
| Med-Qwen2 7B          | 11.9      | 21.4 | 36.2 | 39.4 | 37.7 | 38.6 | 38.1 | 48.0 | 34.2 | 9.5  |  | 31.5    |
| MedGemma 4B           | 0.0       | 26.2 | 38.6 | 44.1 | 32.4 | 33.8 | 32.8 | 25.7 | 27.2 | 35.0 |  | 29.6    |
| Qwen3.5 Flash         | 24.8      |      |      |      |      |      |      |      |      |      |  | 24.8    |
| BioMistral 7B         | 2.6       | 13.4 | 13.0 | 14.9 | 16.5 | 13.3 | 13.5 | 12.2 | 12.7 | 3.6  |  | 11.6    |
| Qwen3 32B             | 0.0       | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |  | 0.0     |
| Qwen3 8B              | 0.0       | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |  | 0.0     |
| KhazarAI Bio 8B       | 0.0       | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |  | 0.0     |
| Average               | 34.3      | 38.1 | 36.3 | 64.4 | 72.0 | 66.3 | 65.9 | 45.6 | 43.4 | 41.8 |  | 69.9    |
|                       | 1         | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   |  | Average |
|                       | Prompt ID |      |      |      |      |      |      |      |      |      |  |         |

Figure 7: Subtask 4 Validation Results (All Models)

# WisPerMed at ArchEHR-QA 2026: Retrieval-Augmented Prompting for Grounded EHR Question Answering

Jan-Henning Büns<sup>1</sup>, Tabea M. G. Pakull<sup>2,1</sup>, Hendrik Damm<sup>1,3</sup>, Bohao Chu<sup>6</sup>,  
Christoph M. Friedrich<sup>1,7</sup>, Felix Nensa<sup>4,5</sup>, Elisabeth Livingstone<sup>2</sup>,  
Peter A. Horn<sup>2</sup>, Norbert Fuhr<sup>6</sup>

<sup>1</sup>Department of Computer Science, University of Applied Sciences and Arts Dortmund

<sup>2</sup>Institute for Transfusion Medicine, University Hospital Essen

<sup>3</sup>Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen

<sup>4</sup>Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen

<sup>5</sup>Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen

<sup>6</sup>Department of Computer Science, University of Duisburg-Essen

<sup>7</sup>Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen

Correspondence: [jan-henning.buens@fh-dortmund.de](mailto:jan-henning.buens@fh-dortmund.de)

## Abstract

ArchEHR-QA is a grounded question-answering (QA) task for electronic health records (EHRs) comprising four subtasks: (1) question rewriting, (2) evidence identification, (3) grounded answer generation, and (4) answer-evidence alignment. In this work, we present a modular pipeline centered on retrieval-augmented generation (RAG). For Subtask 1, RAG few-shot prompting outperformed both PEFT and prompt-only baselines on the development set; however, Claude few-shot proved substantially more robust on the test set, ranking 6th out of 6 participating teams (score: 26.94). For Subtask 2, a union ensemble of open-weight LLMs (GPT-OSS-120B and Qwen3-30B-A3B) achieved a 56.7 micro-F1, rivaling the proprietary Claude Opus 4.6 while demonstrating higher recall (53.6). For Subtask 3, our RAG few-shot approach using Claude Opus 4.5 achieved the 1st place out of 13 participating teams (score: 36.33). Finally, for Subtask 4, a zero-shot Claude Opus 4.6 configuration ranked 2nd out of 16 participating teams (score: 81.3).

**Keywords:** Electronic Health Records, Question Answering, Retrieval-Augmented Generation, Large Language Models, Clinical NLP

## 1. Introduction

Patient portals and electronic messaging systems are fundamentally changing how patients interact with their own health information. As these platforms become widespread, patients increasingly access clinical documentation—discharge summaries, progress notes, laboratory results—and pose questions to their care teams through asynchronous messaging (Johnson et al., 2016). This shift creates a growing demand for automated systems that can assist clinicians in responding to patient inquiries accurately and efficiently. However, building such systems is far from straightforward: patient questions are often verbose, colloquial, and underspecified, while the clinical notes that contain the answers are long, heterogeneous, and densely packed with medical terminology across multiple concurrent problems.

A key requirement for any question-answering (QA) system operating in this clinical setting is *grounding*—the ability to not only generate fluent and medically appropriate responses, but to explicitly link each claim in the response to supporting evidence in the patient’s electronic health record (EHR). Without grounding, even factually correct answers lack the verifiability that clinicians need to trust and adopt automated assistance, and hallucinated content poses a direct risk of patient harm.

This requirement distinguishes clinical QA from open-domain question answering, where fluency and factual accuracy are often sufficient.

The ArchEHR-QA shared task (Soni and Demner-Fushman, 2026b) formalizes these challenges into four complementary subtasks that together constitute a complete grounded QA pipeline: (1) *question interpretation*, which rewrites a verbose patient question into a concise clinician query; (2) *evidence identification*, which selects the minimal set of EHR sentences needed to answer the question; (3) *answer generation*, which produces a grounded response constrained by the clinical evidence; and (4) *answer-evidence alignment*, which maps each answer sentence to its supporting evidence at the sentence level. This decomposition reflects the real-world workflow of clinicians who must first understand the patient’s intent, locate the relevant information, formulate a response, and ensure that every statement is supported.

In this work, we present a modular pipeline for all four subtasks, built around large language models (LLMs) with retrieval-augmented generation (RAG). Our system design is guided by three practical constraints that are central to clinical NLP deployment. First, *format fidelity*: each subtask imposes strict

output constraints (e.g., at most 15 words for question rewriting, at most 75 words for answer generation) that must be enforced reliably. Second, *faithfulness*: generated answers must be grounded in the provided clinical notes without introducing unsupported claims. Third, *reproducibility and data privacy*: clinical data is subject to stringent governance requirements, motivating a systematic comparison between locally deployable open-weight LLMs and proprietary API-based models to quantify the accuracy–privacy trade-off. To support independent development, error analysis, and auditability, each subtask is addressed by a dedicated module that can be evaluated and improved in isolation.

**Contributions.** Our main contributions are:

- A **RAG in-context learning strategy** that dynamically retrieves semantically similar exemplars and enforces task-specific output constraints for question rewriting and answer generation, ranking 1st out of 13 teams on answer generation.
- A **privacy-aware open-LLM ensemble** for evidence identification, combining two open-weight LLMs via a union strategy that achieves competitive performance with a proprietary LLM (56.7 vs. 58.8 micro-F1) while providing higher recall and preserving full data sovereignty.
- Evidence that **patient question context substantially improves answer–evidence alignment**, with a zero-shot prompting approach ranking 2nd out of 16 teams (micro-F1: 81.3).
- A **systematic analysis across all four subtasks**, comparing prompting, RAG, fine-tuning, and preference optimization paradigms, with a detailed examination of how distribution shift degrades retrieval-based methods relative to stronger LLM backbones.

## 2. Related Work

**Grounded QA from clinical text.** Evidence-grounded (QA) has been extensively studied in open-domain settings, but grounding answers in EHR notes introduces additional challenges: long and heterogeneous narratives, multiple concurrent problems, and high consequences of hallucinated content. The ArchEHR-QA shared task formalizes these challenges with explicit evidence annotations and answer–evidence alignments (Soni and Demner-Fushman, 2026b). The dataset is derived from de-identified clinical documentation in the style of MIMIC critical care records, and thus inherits the linguistic variability and documentation conventions typical of real-world EHR notes (Johnson et al., 2016).

**Consumer health question rewriting and summarization.** Subtask 1 is closely related to con-

sumer health question summarization, where verbose patient queries are condensed while preserving intent and clinical content. MeQSum provides expert summaries for such questions and is widely used for training and evaluating medical question summarization systems (Ben Abacha and Demner-Fushman, 2019). Unlike generic rewriting, ArchEHR-QA requires the output to resemble a clinician query to an EHR system, without introducing unsupported content.

**RAG prompting.** RAG (Lewis et al., 2020) and retrieval of in-context demonstrations (e.g., nearest-neighbor prompting) can improve few-shot performance by providing style- and content-aligned exemplars at inference time (Shi et al., 2022). In parallel, parameter-efficient fine-tuning methods such as LoRA reduce the cost of adapting LLMs to domain-specific tasks (Hu et al., 2022), while preference optimization (e.g., DPO) can shift LLM behavior toward desired outputs without explicit reward modeling (Rafailov et al., 2023).

## 3. Data

Each case in the ArchEHR-QA dataset (Soni and Demner-Fushman, 2026a) contains a patient-authored question, a clinician-interpreted question (reference for Subtask 1), a clinical note excerpt segmented into numbered sentences with sentence-level relevance annotations, a clinician-authored answer (reference for Subtask 3), and answer–evidence alignments (reference for Subtask 4) (Soni and Demner-Fushman, 2026b). The dataset is distributed via PhysioNet and is derived from de-identified EHR documentation (MIMIC-style) (Johnson et al., 2016).

## 4. System Overview

Our system follows a modular pipeline aligned with the subtasks of ArchEHR-QA (Organizers, 2026).

**Subtask 1: Question Interpretation.** Given a patient-authored question, the system must produce a single clinician-interpreted question that captures the core information need in at most 15 space-separated words, ends with a question mark, and does not introduce facts absent from the input. We address this with RAG few-shot prompting, dynamically selecting style-matched exemplars from the training set to guide format and clinical register.

**Subtask 2: Evidence Identification.** Given a patient question, a clinician question (gold or system-generated), and a numbered clinical note excerpt,

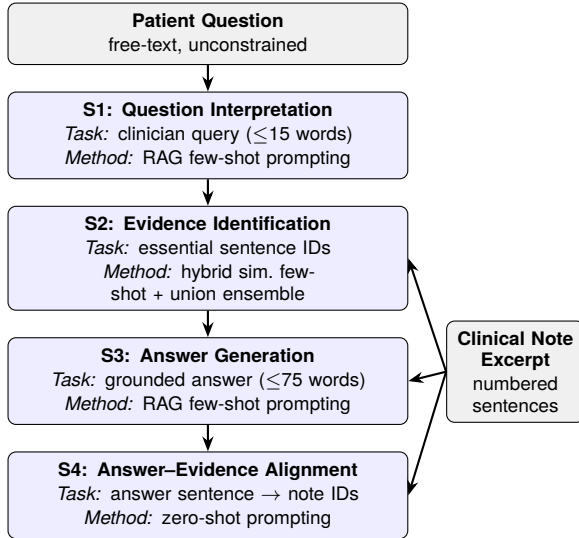


Figure 1: Modular pipeline for ArchEHR-QA. Each box states the task output constraint and the method used.

the system must identify the minimal set of sentence IDs that sufficiently support answering the question. We approach this as a direct selection task using hybrid similarity-based few-shot in-context learning, and combine predictions from two open-weight LLMs via a union ensemble that exploits their complementary precision–recall profiles.

**Subtask 3: Answer Generation.** Given a patient question, a clinician question, and a clinical note excerpt, the system must generate a grounded answer in professional register, limited to at most 75 words and avoiding speculation where evidence is partial. We generate answers conditioned on the full note excerpt with evidence emphasis, using RAG few-shot prompting to supply style- and content-aligned demonstrations at inference time.

**Subtask 4: Answer–Evidence Alignment.** Given a patient question, a numbered clinical note excerpt, and a numbered answer text, the system must map each answer sentence to the set of note sentences that support it, producing many-to-many alignments. We adopt a zero-shot prompting strategy, instructing the LLM to identify supporting evidence sentences for each answer sentence individually.

## 5. Selection of LLMs

We primarily use open-weight instruction-tuned LLMs for local inference and reproducibility: Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) and OpenAI’s open-weight gpt-oss-120b (OpenAI, 2025), served via vLLM (Kwon et al., 2023). For

all subtasks comparisons (patient question only), we also report closed-LLM baselines using Claude Opus 4.5 (Anthropic, 2025) and Claude Opus 4.6 (Anthropic, 2026).

## 6. Subtask 1: Question Interpretation

### 6.1. Prompting and Constraint Enforcement

All Subtask 1 approaches share a common instruction that explicitly encodes the constraints. We apply deterministic post-processing: (i) strip artifacts (e.g., bullets, quotation marks, prefixes such as “Answer:”), (ii) enforce a trailing “?” if missing, and (iii) enforce the 15-word limit by whitespace tokenization and truncation (consistent with the official evaluation policy).

Decoding uses temperature 0.2, top- $p$  0.9, max 64 new tokens, and two random seeds.

### 6.2. Prompting Strategies

**Zero-shot prompting.** We provide only the constraint instruction and the patient question.

**Few-shot prompting.** We prepend four exemplars (patient question → clinician question) to induce the desired query style.

**RAG few-shot prompting.** We implement RAG in-context learning by selecting exemplars dynamically for each test input. We embed candidate exemplars with SentenceTransformers (Reimers and Gurevych, 2019) (all-mpnet-base-v2; MPNet backbone (Song et al., 2020)) and build a FAISS inner-product index (Johnson et al., 2017). At inference time, the patient question retrieves top- $k$  similar training inputs; we inject the top exemplars as demonstrations and then generate the clinician question. This approach is closely related to nearest-neighbor prompting (Shi et al., 2022), but adapted to a clinical rewriting format with explicit length constraints.

**Clinical retrieval variant.** We also test a clinical embedding variant using BioBERT (Lee et al., 2020) to compute dense sentence representations and (optionally) a cross-encoder reranker (details in the appendix).

**Closed-LLM baseline.** We run Claude Opus 4.5 (Anthropic, 2025) with the same zero-shot and few-shot prompts for a closed vs. open comparison on Subtask 1.



### 6.3. Supervised Fine-Tuning and Preference Optimization

**SFT.** We fine-tune Qwen3-30B-A3B-Instruct-2507 and gpt-oss-120b with LoRA adapters (Hu et al., 2022) using chat-formatted inputs.

**DPO.** We apply DPO (Rafailov et al., 2023) on top of the SFT LLM by sampling candidates per prompt and constructing preference pairs using ROUGE-L as a heuristic selector. (We include this as an exploratory alignment method; see Discussion for limitations.)

## 7. Subtask 2: Evidence Identification

### 7.1. Hybrid Similarity-Based Few-Shot Prompting

**Example selection.** We select  $k=5$  few-shot exemplars per target case using a hybrid similarity measure that combines semantic and lexical signals:

$$\text{sim}(q, q_i) = \alpha \cdot \cos(\mathbf{e}_q, \mathbf{e}_{q_i}) + (1 - \alpha) \cdot \text{RL}(q, q_i) \quad (1)$$

where  $\mathbf{e}$  denotes sentence embeddings from all-MiniLM-L6-v2 (Reimers and Gurevych, 2019), RL is the ROUGE-L F-measure (Lin, 2004), and  $\alpha=0.7$ . The combination favors semantically similar cases while also rewarding surface-level overlap in clinical terminology, which is important for matching cases within the same clinical specialty. Examples are drawn from the development set, excluding the current target case.

**Prompt construction.** The prompt consists of: (i) a system instruction defining the evidence selection role, (ii) a task description with five criteria for essential evidence (sentences that directly address the question, contain critical clinical findings, provide necessary diagnostic or treatment context, are indispensable for a complete answer, or include key lab values/medications/procedures), (iii) the  $k$  most similar few-shot examples, each presenting the clinician question, numbered note sentences, and gold essential sentence IDs, and (iv) the target case without the answer. The LLM is instructed to respond in a structured format: `Answer: The relevant sentence IDs are: <comma-separated IDs>.`

**Decoding and output parsing.** All LLMs use greedy decoding (temperature 0.0) for deterministic, reproducible output. A regex-based parser extracts sentence IDs from the structured output, handling variations such as JSON lists, free-text ID mentions, and whitespace differences.

**Union ensemble.** We combine predictions from GPT-OSS-120B and Qwen3-30B-A3B using a union (logical OR) strategy: a sentence ID is included in the final prediction if either LLM selects it. This design exploits the complementary precision–recall profiles observed during development: GPT-OSS-120B tends toward conservative, high-precision predictions (strict micro-P 71.2% on dev), while Qwen3-30B-A3B favors broader selection with higher recall (strict micro-R 63.6% on dev). Both LLMs are served simultaneously and their prediction sets are merged and sorted by sentence ID.

**Closed-LLM baseline.** To contextualize the open-LLM ensemble, we run the same hybrid similarity-based 5-shot prompting pipeline (Section 7.1) with Claude Opus 4.6 (Anthropic, 2026) via the Anthropic API. This provides a single-LLM proprietary baseline that uses the identical prompt template, example selection strategy, and output parsing, isolating the effect of LLM choice from all other pipeline components. Comparing the ensemble against this baseline quantifies whether combining two smaller open-weight LLMs can match or exceed a state-of-the-art closed LLM under controlled conditions.

## 8. Subtask 3: Answer Generation

### 8.1. Shared Prompt and Post-Processing

All Subtask 3 approaches share a system prompt instructing the LLM to act as a clinical QA assistant with explicit constraints:  $\leq 75$  words, use only facts from the clinical note excerpt, maintain professional register, and avoid speculation. The user prompt structures the input as three labeled sections (patient question, clinician question, clinical note excerpt).

All outputs undergo deterministic post-processing: (i) removal of `<think>` blocks (common in reasoning-enabled LLMs), (ii) stripping of quotation marks, bullets, and LLM-specific prefixes, (iii) medical term normalization, and (iv) truncation to 75 words with proper sentence-ending punctuation.

### 8.2. Prompting Strategies

**Zero-shot prompting.** The LLM receives only the system instruction and the case (no exemplars). We evaluate three backbones: Qwen3-30B-A3B, GPT-OSS-120B, and Claude Opus 4.5. Decoding uses temperature 0.3, top- $p$  0.9. For local LLMs, two candidates are generated with different seeds and reranked; Claude uses a single seed.

**Few-shot prompting.** Four fixed exemplars (patient question, clinician question, note excerpt, gold answer) are prepended as user–assistant turn pairs. Exemplars are randomly sampled from the training split (seed 13).

**RAG few-shot.** Instead of fixed exemplars, we dynamically retrieve the four most relevant training cases per test input from a FAISS inner-product index built with `all-mpnet-base-v2` embeddings (§6.2). The query concatenates the patient and clinician questions; top-4 neighbors are retrieved and used as demonstrations.

**Closed-LLM baseline.** We run Claude Opus 4.5 (Anthropic, 2025) with the same zero-shot and few-shot prompts for a closed vs. open comparison.

### 8.3. Supervised Fine-Tuning and Preference Optimization

**SFT.** We fine-tune Qwen3-30B-A3B-Instruct-2507 with QLoRA (4-bit NF4) on training cases formatted as chat interactions. Two configurations are explored: (i) standard (LoRA  $r=8$ ,  $\alpha=16$ , 5 epochs, max length 1024; best checkpoint at step 35) and (ii) extended (LoRA  $r=16$ ,  $\alpha=32$ , 10 epochs, max length 2048; best checkpoint at step 70). Both use learning rate  $2 \times 10^{-5}$ , effective batch size 16, early stopping on validation loss, and AdamW.

**DPO.** Starting from the base LLM, we generate multiple candidates per training case at varied temperatures (0.2–0.4), score them against reference answers with ROUGE-L, and select the best/worst as chosen/rejected pairs. DPO training uses  $\beta=0.1$ , LoRA  $r=8$  (standard) or  $r=16$  (extended), and learning rates of  $5 \times 10^{-6}$  and  $1 \times 10^{-6}$  respectively.

## 9. Subtask 4: Evidence Alignment

### 9.1. Prompting Strategy

**LLM selection.** We evaluated three state-of-the-art closed-source LLMs via API on the development set: GPT-5.1 (OpenAI, 2025), GPT-5.2 (OpenAI, 2025), and Claude Opus 4.6 (Anthropic, 2026). Based on development set performance, Claude Opus 4.6 achieved the best results and was selected for final evaluation on the test set.

**Zero-shot prompting.** We adopt a zero-shot prompting strategy without any task-specific demonstrations. For each answer sentence  $a_i \in \mathcal{A}$ , we construct a prompt that includes the patient question  $q$ , the full clinical note excerpt  $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$ , and the complete answer  $\mathcal{A} =$

$\{a_1, a_2, \dots, a_m\}$  as context, and instruct the LLM to identify the supporting evidence sentences from  $\mathcal{S}$  for the given  $a_i$ . Formally, the LLM is queried as  $f(a_i | q, \mathcal{S}, \mathcal{A}) \rightarrow \mathcal{S}' \subseteq \mathcal{S}$ , where  $\mathcal{S}'$  is the predicted set of supporting sentences to create a mapping  $f : \mathcal{A} \times \mathcal{S} \rightarrow \{0, 1\}$  that aligns each answer sentence to its supporting sentence(s) in the clinical note excerpt.

## 10. Experimental Setup

**Metrics.** We follow the official evaluation scripts (Organizers, 2026). For Subtask 1, we report ROUGE-L, BERTScore, AlignScore, and MEDCON; the primary score is  $(\text{ROUGE-L} + \text{BERTScore})/2$  (Lin, 2004; Zhang et al., 2020; Laban et al., 2022). MEDCON computes clinical concept overlap using UMLS concept extraction (QuickUMLS) (Bodenreider, 2004; Soldaini and Goharian, 2016). For Subtask 2, we report strict and lenient micro/macro precision, recall, and F1. Under strict evaluation, only essential sentences count as gold positives; under lenient evaluation, supplementary sentences additionally do not incur false-positive penalties. For Subtask 3, the official overall score aggregates BLEU, ROUGE-1/2/L/Lsum, SARI, BERTScore, AlignScore, and MEDCON. During development we used a primary score of  $(\text{ROUGE-L} + \text{BERTScore})/2$  and BLEU as a secondary metric. For Subtask 4, we report micro-averaged Precision, Recall, and F1 over all predicted alignment links as the primary evaluation metrics. In addition, we report macro-averaged Precision, Recall, and F1 as supplementary analysis.

**Implementation notes.** We use two random seeds for generation-based runs (Subtask 1) and perform deterministic post-processing to meet format constraints. Subtask 2 uses greedy decoding (temperature 0.0) for deterministic output. Open-weight LLMs are served via vLLM on NVIDIA H100 GPUs; Claude models are accessed via the Anthropic API. Training and adapter configurations are documented in Appendix A.

## 11. Results

### 11.1. Subtask 1 Results

We report validation results on 100 released cases (IDs 21–120) using the official scorer. Table 1 summarizes our main Subtask 1 approaches. On the development set, RAG few-shot prompting achieves the highest primary score (31.05) and overall score (25.61), followed by the clinical retrieval variant (25.41) and Claude few-shot (25.27). DPO achieves the highest AlignScore (18.37) and

| Approach                | Primary      | R-L          | BERT         | Align        | MEDCON       | Overall      |
|-------------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| RAG few-shot            | <b>31.05</b> | <b>23.63</b> | 38.46        | 14.91        | 25.45        | <b>25.61</b> |
| RAG few-shot (clinical) | 30.78        | 23.34        | 38.22        | 15.18        | 24.92        | 25.41        |
| Claude few-shot         | 30.84        | 22.84        | <b>38.84</b> | 14.90        | 24.51        | 25.27        |
| Qwen SFT                | 27.61        | 18.79        | 36.44        | 17.76        | 24.11        | 24.27        |
| Qwen DPO (full)         | 24.53        | 16.44        | 32.61        | <b>18.37</b> | <b>26.58</b> | 23.50        |

Table 1: Validation scores for Subtask 1 on cases 21–120 (100 cases). Primary = (ROUGE-L + BERTScore)/2 on 0–100 scale. Overall = official mean of AlignScore, ROUGE-L, MEDCON, and BERTScore. MEDCON uses UMLS concept extraction (QuickUMLS) with UMLS knowledge sources.

| Approach  | R-L          | BERT         | Align        | MC           | OA           |
|-----------|--------------|--------------|--------------|--------------|--------------|
| Claude fs | <b>21.76</b> | <b>38.01</b> | <b>21.71</b> | <b>26.28</b> | <b>26.94</b> |
| RAG fs    | 18.46        | 33.49        | 14.39        | 18.99        | 21.33        |
| Qwen DPO  | 10.68        | 20.55        | 16.77        | 20.83        | 17.21        |

Table 2: Codabench Subtask 1 test scores (cases 121–167). MC = MEDCON, OA = overall score.

| # | Team                 | OA          |
|---|----------------------|-------------|
| 1 | HealthNLP_Retrievers | 31.2        |
| 2 | KPSCMI               | 30.8        |
| 3 | OptiMed              | 29.9        |
| 4 | Neural               | 28.9        |
| 5 | Yale-DM-Lab          | 27.1        |
| 6 | <b>Ours</b>          | <b>26.9</b> |

Table 3: Subtask 1 Codabench leaderboard (6/13 teams). OA = official overall score.

MEDCON (26.58) but the lowest primary and overall scores.

**Test set (cases 121–167).** The official test labels are withheld; scores are returned via Codabench. Our three Subtask 1 submissions are Claude few-shot, RAG few-shot, and Qwen DPO (full-train). Table 2 lists the official test scores; Table 3 provides leaderboard context.

On the test set, the ranking reverses substantially: Claude few-shot achieves the best overall score among our submissions (26.94), outperforming RAG few-shot (21.33) by 5.6 points. Claude few-shot leads on all individual metrics, with particularly large margins on BERTScore (38.01 vs. 33.49) and MEDCON (26.28 vs. 18.99). Qwen DPO trails further (17.21). Our best submission (Claude few-shot) places 6th on the official leaderboard, 4.3 points behind the top-ranked system (HealthNLP\_Retrievers, 31.2).

## 11.2. Subtask 2 Results

We submitted three configurations to the official Codabench leaderboard: (1) Claude Opus 4.6 with 5-shot prompting, (2) the union ensemble of GPT-OSS-120B and Qwen3-30B-A3B with 5-shot

prompting, and (3) the same ensemble in a zero-shot setting (no in-context examples). Table 4 reports the official test scores.

Claude Opus 4.6 achieves the highest strict micro-F1 among our submissions (58.8), placing 10th on the leaderboard out of 14 teams. The open-LLM ensemble with few-shot prompting reaches 56.7 (12th), trailing Opus by only 2.1 F1 points while achieving higher recall (53.6 vs. 52.3). Removing few-shot examples (zero-shot ensemble) drops performance substantially to 51.4, confirming the value of in-context demonstrations for calibrating evidence selection.

All three runs exhibit notably high lenient precision (up to 84.2% for Opus), indicating that sentences classified as false positives under strict evaluation are predominantly supplementary rather than not-relevant—the LLMs select clinically pertinent sentences even when they do not meet the strict essential threshold.

## 11.3. Subtask 3 Results

We evaluated 10 configurations on the 20-case development set, covering four paradigms (zero-shot, few-shot, RAG few-shot, SFT/DPO), and submitted three systems to the official test phase. Table 5 summarizes development results for representative approaches; Table 6 reports official test scores.

**Development set results.** RAG few-shot with Claude Opus 4.5 achieves the highest primary score (60.5), BLEU (12.7), and MEDCON (54.6) on the 20-case development set (Table 5) and outperforms the next-best configuration (RAG few-shot Qwen) by 2.4 primary points. Dynamically retrieved exemplars consistently outperform fixed few-shot demonstrations across both LLM backbones (Claude: 60.5 vs. 57.8; Qwen: 58.1 vs. 55.1).

**Effect of fine-tuning.** Extended SFT training improves over the standard configuration by 2.2 primary points (57.2 vs. 55.0). Both SFT variants trail RAG few-shot Qwen (58.1). DPO does not improve over SFT (53.3 vs. 55.0).

| Run                                       | OA          | Strict Micro |             |             |             | Lenient Micro |             |             |             |
|-------------------------------------------|-------------|--------------|-------------|-------------|-------------|---------------|-------------|-------------|-------------|
|                                           |             | P            | R           | F1          | Macro F1    | P             | R           | F1          | Macro F1    |
| Opus 4.6 (5-shot)                         | <b>58.8</b> | <b>67.1</b>  | 52.3        | <b>58.8</b> | <b>58.1</b> | <b>84.2</b>   | 52.3        | <b>64.5</b> | <b>63.6</b> |
| Ensemble (5-shot)                         | 56.7        | 60.2         | <b>53.6</b> | 56.7        | 55.1        | 74.7          | <b>53.6</b> | 62.4        | 60.4        |
| Ensemble (0-shot)                         | 51.4        | 61.8         | 44.0        | 51.4        | 50.3        | 76.7          | 44.0        | 55.9        | 54.6        |
| <i>Leaderboard context (top-3 teams):</i> |             |              |             |             |             |               |             |             |             |
| 1st (Neural)                              | 63.7        | 60.2         | 67.6        | 63.7        | 64.8        | 77.8          | 67.6        | 72.4        | 73.1        |
| 2nd (OptiMed)                             | 63.2        | 56.7         | 71.3        | 63.2        | 64.1        | 73.1          | 71.3        | 72.2        | 72.9        |
| 3rd (UIC-AIHealth4All)                    | 62.9        | 59.3         | 67.0        | 62.9        | 63.0        | 74.3          | 67.0        | 70.4        | 70.5        |

Table 4: Subtask 2 evidence identification results on the official Codabench test set (14 participating teams). OA = Overall (Strict Micro F1), the official ranking metric. Ensemble = union of GPT-OSS-120B and Qwen3-30B-A3B. Macro F1 = strict macro F1 for our runs; shown for comparison. Best values among our submissions in bold.

| Approach         | Primary     | BLEU        | MC          |
|------------------|-------------|-------------|-------------|
| RAG fs Claude    | <b>60.5</b> | <b>12.7</b> | <b>54.6</b> |
| RAG fs Qwen      | 58.1        | 9.1         | 46.7        |
| Few-shot Claude  | 57.8        | 9.6         | 46.9        |
| SFT-long Qwen    | 57.2        | 8.3         | 39.0        |
| Few-shot Qwen    | 55.1        | 6.3         | 42.4        |
| SFT Qwen         | 55.0        | 4.5         | 34.5        |
| Zero-shot Claude | 54.1        | 6.3         | 43.1        |
| DPO Qwen         | 53.3        | 4.2         | 35.7        |
| Zero-shot Qwen   | 53.0        | 4.0         | 35.6        |
| RAG fs OSS       | 52.3        | 5.8         | 46.0        |
| Few-shot OSS     | 52.3        | 5.7         | 43.2        |
| Zero-shot OSS    | 51.4        | 4.9         | 41.3        |

Table 5: Subtask 3 development results (20 cases; selected configurations). Primary = (ROUGE-L + BERTScore)/2  $\times$  100. MC = MEDCON.

**Backbone comparison.** Under identical prompting, a consistent ranking emerges across backbones. With RAG few-shot, Claude Opus 4.5 outperforms Qwen3-30B-A3B by 2.4 primary points and 7.9 MEDCON points; under zero-shot, the MEDCON gap widens further (43.1 vs. 35.6). GPT-OSS-120B consistently trails both Claude and Qwen and records systematically lower primary scores.

**Test set results and leaderboard placement.** RAG few-shot Claude achieves an overall score of 36.33 on the official test set (Table 6) and ranks **first** out of 13 participating teams (Table 7). The margin over the second-ranked team (TAMU-NLP-Lab, 36.2) is narrow ( $\Delta=0.1$ ). RAG few-shot Claude leads on all individual metrics except SARI, where SFT-long scores highest (59.31 vs. 58.60).

**RAG gain over zero-shot.** On the test set, RAG few-shot Claude outperforms zero-shot Claude by 3.5 overall points (36.33 vs. 32.83). The largest margins appear on BLEU (9.94 vs. 5.28) and

BERTScore (46.79 vs. 41.76). The MEDCON gap is smaller (43.14 vs. 41.47).

#### 11.4. Subtask 4 Results

We first present a comparative analysis of the LLMs described in Section 9.1 on the development set under both the w/ and w/o patient question settings, followed by the evaluation of the best-performing LLM on the test set.

**Comparison across LLMs.** As shown in the lower half of Table 8, Claude Opus 4.6 achieves the best overall performance across all metrics, attaining a micro F1 of 0.975 and a macro F1 of 0.973 under the w/ patient question setting, substantially outperforming both GPT variants. GPT-5.2 consistently ranks second, with micro F1 scores of 0.868 and 0.878 in the two settings, while GPT-5.1 records the lowest scores across all metrics. Notably, all three LLMs achieve consistently high recall, suggesting that LLMs are generally capable of retrieving relevant evidence sentences; however, precision varies more markedly across LLMs, with Opus 4.6 demonstrating a considerably stronger ability to suppress spurious alignments.

**Effect of patient question.** As shown in the upper half of Table 8, incorporating the patient question  $q$  into the input context consistently improves performance across all LLMs and metrics. The most substantial gains are observed for Claude Opus 4.6, where micro Precision increases from 0.783 to 0.958 and micro F1 improves from 0.875 to 0.975, indicating that the patient question provides critical contextual signal for disambiguating relevant evidence sentences. GPT-5.1 and GPT-5.2 also benefit from the inclusion of  $q$ , with moderate gains in precision and F1 while recall remains relatively stable. These results highlight the importance of the patient question as contextual grounding, particularly for more capable LLMs that can better exploit

| Run                    | OA           | Align        | BERT         | BLEU        | MC           | R-L          | SARI         |
|------------------------|--------------|--------------|--------------|-------------|--------------|--------------|--------------|
| RAG fs Claude          | <b>36.33</b> | <b>31.71</b> | <b>46.79</b> | <b>9.94</b> | <b>43.14</b> | <b>27.81</b> | 58.60        |
| SFT-long (Qwen, local) | 34.16        | 29.86        | 43.56        | 8.36        | 37.53        | 26.30        | <b>59.31</b> |
| Zero-shot Claude       | 32.83        | 28.08        | 41.76        | 5.28        | 41.47        | 23.12        | 57.28        |

Table 6: Subtask 3 official test results. OA = overall score (composite of all metrics). MC = MEDCON. Best values in bold.

| #                 | Team                 | OA          |
|-------------------|----------------------|-------------|
| 1                 | <b>Ours</b>          | <b>36.3</b> |
| 2                 | TAMU-NLP-Lab         | 36.2        |
| 3                 | BIT.UA-AAUBS         | 35.6        |
| 4                 | Neural               | 35.2        |
| 5                 | HealthNLP_Retrievers | 34.6        |
| Median (13 teams) |                      | 32.3        |

Table 7: Subtask 3 Codabench leaderboard (top-5 of 13 teams). OA = official overall score.

| Run                  | $P_{mi}$     | $R_{mi}$     | $F1_{mi}$    | $P_{ma}$     | $R_{ma}$     | $F1_{ma}$    |
|----------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| W/O Patient Question |              |              |              |              |              |              |
| GPT-5.1              | 0.735        | 0.884        | 0.803        | 0.752        | 0.887        | 0.797        |
| GPT-5.2              | 0.799        | <u>0.949</u> | 0.868        | 0.847        | 0.947        | 0.882        |
| Opus-4.6             | 0.783        | <b>0.993</b> | 0.875        | 0.812        | <b>0.993</b> | 0.876        |
| W/ Patient Question  |              |              |              |              |              |              |
| GPT-5.1              | 0.774        | 0.891        | 0.835        | 0.811        | 0.894        | 0.835        |
| GPT-5.2              | <u>0.827</u> | 0.935        | <u>0.878</u> | <u>0.878</u> | 0.937        | <u>0.896</u> |
| Opus-4.6             | <b>0.958</b> | <b>0.993</b> | <b>0.975</b> | <b>0.962</b> | <u>0.990</u> | <b>0.973</b> |

Table 8: Subtask 4 answer–evidence alignment results with and without the patient question  $q$  as input context. Subscripts  $_{mi}$  and  $_{ma}$  denote micro- and macro-averaged metrics, respectively. **Bold** indicates the best score and underline indicates the second-best score in each column.

the additional information.

**Test Set Results.** We submitted the predictions of Claude Opus 4.6 with patient question included for the final evaluation on the test set. The system achieves a micro F1 of 81.27 and a macro F1 of 82.09, with a micro Precision of 86.93, micro Recall of 76.31, macro Precision of 87.54, and macro Recall of 79.41. Compared to the development set results, there is a noticeable drop in both precision and recall, which is expected due to the distribution shift between the development and test sets. Nevertheless, the system maintains relatively strong precision, suggesting that the predicted alignment links are generally reliable.

## 12. Discussion

**Subtask 1 performance trends and test-time shift.** On cases 21–120, RAG few-shot prompting is consistently strongest, with the clinical retrieval variant close behind. The Claude few-shot baseline is competitive and outperforms open baselines on several metrics. SFT improves over prompt-only baselines but trails RAG on the official overall score. DPO increases AlignScore and MEDCON but does not translate into higher overall performance. On the test set (cases 121–167), the ranking reverses substantially: Claude few-shot leads by 5.6 overall points (26.94 vs. 21.33), whereas RAG few-shot led on development by only 0.34 points. This highlights a significant distribution shift between the released labeled cases and the hidden test set. RAG few-shot retrieves exemplars from the training pool using semantic similarity; when test inputs diverge from this pool, the retrieved demonstrations may be poorly matched, amplifying template-like phrasing and reducing lexical as well as semantic overlap with the references. In contrast, the Claude backbone’s stronger language modeling generalizes more robustly without reliance on exemplar quality. The large drop in RAG few-shot MEDCON (25.45 on dev vs. 18.99 on test) suggests that retrieved exemplars may also bias clinical terminology toward training-set conventions that do not transfer. The DPO configuration degrades further on the test set (17.21), confirming that ROUGE-L-based preference optimization does not align with the official multi-metric objective, particularly under distribution shift.

**Subtask 2: Open-LLM ensemble vs. proprietary baseline.** Claude Opus 4.6 achieves the highest strict micro-F1 among our submissions (58.8), driven by substantially higher precision (67.1 vs. 60.2). However, the union ensemble of two open-weight LLMs achieves higher recall (53.6 vs. 52.3) and remains competitive on F1 (56.7 vs. 58.8). This demonstrates that complementary open LLMs—one exhibiting higher precision (GPT-OSS-120B) and one exhibiting higher recall (Qwen3-30B-A3B) in our experiments—can approach proprietary-LLM performance while preserving full data sovereignty, which is particularly relevant in clinical settings where note text may be subject to data use agreements. A gap of approximately 5 F1 points to the



top-ranked system (Neural, 63.7) suggests room for improvement, potentially through threshold tuning, additional ensemble members, or fine-tuning on evidence labels.

**Impact of few-shot exemplars on evidence selection.** Comparing the ensemble in few-shot and zero-shot settings reveals a consistent benefit of in-context demonstrations: 5-shot prompting improves strict micro-F1 by 5.3 points (56.7 vs. 51.4) and recall by 9.6 points (53.6 vs. 44.0). The hybrid similarity-based selection (Eq. 1) retrieves cases from similar clinical specialties, providing the LLM with domain-appropriate evidence patterns. This confirms that example selection substantially calibrates the LLM’s evidence threshold, reducing both under-selection (zero-shot conservatism) and over-selection.

**Strict vs. lenient evaluation gap.** All Subtask 2 runs exhibit a large gap between strict and lenient precision (e.g., 67.1 vs. 84.2 for Opus), indicating that the majority of strict false positives are supplementary sentences rather than not-relevant ones. From a clinical utility perspective, this is encouraging: the LLMs over-predict somewhat but predominantly select sentences that are still clinically pertinent.

**Subtask 3: RAG gain over instruction-only prompting.** Our RAG few-shot Claude submission ranks first out of 13 teams on the official leaderboard (36.3), narrowly ahead of TAMU-NLP-Lab (36.2) and BIT.UA-AAUBS (35.6). On the test set, RAG few-shot Claude outperforms zero-shot Claude by 3.5 overall points, with the largest gaps on BLEU and BERTScore. The MEDCON gap is notably smaller, which indicates that zero-shot Claude already captures relevant clinical concepts but produces surface-level phrasing that deviates from reference conventions. Dynamically retrieved exemplars therefore primarily guide answer style and structure rather than clinical content.

**Local fine-tuning as a privacy-preserving alternative.** The locally fine-tuned SFT-long submission achieves 94% of the top score without API access and thus represents a viable option for locally hosted clinical QA under data-privacy constraints. DPO did not improve over SFT on the development set, likely because ROUGE-L provides too coarse a preference signal to capture the nuanced quality that the multi-metric evaluation rewards. This finding aligns with the Subtask 1 observation that ROUGE-L-based preference optimization does not transfer reliably to composite evaluation objectives.

**LLM quality vs. LLM size.** Under identical prompting, Claude Opus 4.5 consistently outperforms both open-weight LLMs. More notably, the Qwen3-30B-A3B MoE architecture outperforms GPT-OSS-120B despite fewer active parameters. GPT-OSS records systematically lower primary scores, which points to weaker semantic alignment with clinical reference answers. Raw parameter count alone is therefore a poor predictor of clinical text generation quality, and architecture choice as well as training data composition likely play a more decisive role.

**Implications for downstream grounding.** Subtask 1 errors propagate: an overly generic clinician query may reduce evidence recall (Subtask 2), while an overly specific query may bias selection toward irrelevant details. We therefore recommend that Subtask 2 systems (i) optionally incorporate both patient and clinician questions as queries and (ii) include conservative fallbacks for ambiguous interpretation.

## 13. Conclusion

We described a modular system for four ArchEHR-QA subtasks, centered on RAG prompting and constraint-aware generation. For Subtask 1, RAG few-shot prompting yields the strongest development performance, but Claude few-shot proves substantially more robust on the hidden test set (26.94 vs. 21.33 overall), indicating that retrieval-based exemplar selection is sensitive to distribution shift while a stronger LLM backbone generalizes more reliably. For Subtask 2, Claude Opus 4.6 achieves the best overall F1 (58.8), while a union ensemble of two open-weight LLMs attains competitive performance (56.7) with higher recall, offering a viable open-source alternative under data-privacy constraints. For Subtask 3, we systematically compared 13 configurations across prompting, RAG, SFT, and DPO paradigms. RAG few-shot with Claude Opus 4.5 achieves the highest official test score (36.33), ranking first out of 13 teams on the Codabench leaderboard. RAG provides the largest individual gain; local SFT with QLoRA achieves 94% of the top score while preserving full data sovereignty. For Subtask 4, we employ Claude Opus 4.6 (API) in a prompt-only, zero-shot configuration to identify and extract the EHR sentences that support the given response sentence. This approach achieves an official micro-F1 score of 81.3, ranking 2nd out of 16 participating teams.

## 14. Limitations

**Automatic metrics.** Our evaluation relies on automatic metrics (ROUGE, BERTScore, AlignScore, MEDCON for Subtask 1; precision/recall/F1 for Subtask 2) and lacks clinician review; correlations with real clinical utility may be imperfect.

**Data scale and shift.** The released labeled split used for Subtask 1 development is limited, which can amplify variance. For Subtask 2, the development set of 20 cases also serves as the few-shot example pool, which may limit example diversity. For Subtask 3, the same 20-case dev set yields wide metric confidence intervals. Distribution shift between released data and the hidden test set may substantially affect retrieval-based and preference-optimized systems.

**Closed-LLM dependence.** While we report a strong closed-LLM baseline for both Subtask 1 (Claude Opus 4.5) and Subtask 2 (Claude Opus 4.6), reproducibility and cost considerations may limit their use in practice. The open-LLM ensemble addresses this partially but still relies on large-scale infrastructure for serving 120B-parameter LLMs.

## 15. Ethics Statement

This work uses de-identified clinical data distributed via PhysioNet under an approved data use agreement. All experiments were conducted on data that had undergone the MIMIC de-identification pipeline; no attempt was made to re-identify patients, and no individual patient data is reported in this paper.

Our system is designed as a research prototype for a shared task and is not intended for direct clinical deployment. Automatically generated answers to patient questions carry a risk of factual errors, hallucinated clinical content, or misleading omissions that could cause patient harm if acted upon without clinician review. We therefore emphasize that any downstream application of these methods must include clinician-in-the-loop verification before information is communicated to patients.

The use of proprietary LLM APIs (Claude, GPT) for processing clinical text raises data governance concerns, as patient data may traverse third-party infrastructure. We address this partially by evaluating open-weight alternatives that can be hosted locally, preserving data sovereignty. We encourage future work to prioritize locally deployable models for clinical NLP applications where data use agreements restrict external data transfer.

The computational cost of serving large language models (up to 120B parameters on H100

GPUs) has environmental implications. We report all model sizes and hardware configurations to support reproducibility and enable informed comparisons of the accuracy–efficiency trade-off.

## 16. Bibliographical References

- Anthropic. 2025. Claude opus 4.5 system card. <https://www.anthropic.com/claude-opus-4-5-system-card>. Accessed: 2026-01-23.
- Anthropic. 2026. Claude opus 4.6 system card. <https://www.anthropic.com/claude-opus-4-6-system-card>. Accessed: 2026-03-02.
- Anthropic. 2026. [Introducing Claude Opus 4.6](#). Technical report, Anthropic.
- Asma Ben Abacha and Dina Demner-Fushman. 2019. [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.
- Olivier Bodenreider. 2004. [The unified medical language system \(UMLS\): integrating biomedical terminology](#). *Nucleic Acids Research*, 32(suppl 1):D267–D270.
- Edward 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*.
- Hugging Face. 2024. Trl: Transformer reinforcement learning. <https://huggingface.co/docs/trl/>. Accessed: 2026-01-23.
- 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.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. [Billion-scale similarity search with GPUs](#). *arXiv*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

- Philippe Laban, Chien-Sheng Xu, Rahul Aralikkatte, Avinash Thawani, Alexander R. Fabbri, Trevor Cohn, and Marti A. Hearst. 2022. [AlignScore: Evaluating factual consistency with a unified entailment framework](#). *arXiv*.
- 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, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*.
- OpenAI. 2025. [GPT-5.1: A smarter, more conversational ChatGPT](#). Technical report, OpenAI.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *arXiv*.
- OpenAI. 2025. [Introducing GPT-5.2](#). Technical report, OpenAI.
- ArchEHR-QA Organizers. 2026. ArchEHR-QA @ LREC 2026 shared task: Grounded question answering from electronic health records. <https://archehr-qa.github.io/>. Accessed: 2026-01-23.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). *arXiv*.
- 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*.
- Weijia Shi, Ruiqi Gao, Yue Zhang, Yiming Chen, Jie Zhou, Xuanjing Huang, and Furu Wei. 2022. [kNN-prompt: Nearest neighbor zero-shot inference](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.
- Luca Soldaini and Nazli Goharian. 2016. [Quick-UMLS: a fast, unsupervised approach for medical concept extraction](#). In *Proceedings of the 2016 Workshop on Biomedical Natural Language Processing*.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. [MPNet: Masked and permuted pre-training for language understanding](#). In *Advances in Neural Information Processing Systems*.
- Sarvesh Soni and Dina Demner-Fushman. 2026a. [A dataset for addressing patient’s information needs related to clinical course of hospitalization](#). *Scientific Data*.
- 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.
- 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*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). *International Conference on Learning Representations*.

## 17. Language Resource References

- Asma Ben Abacha and Dina Demner-Fushman. 2019. *MeQSum: A Corpus of Medical Question Summarization*. U.S. National Library of Medicine. PID <https://github.com/abachaa/MeQSum>.
- 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 Clinical Database*. PhysioNet, 1.4. PID <https://doi.org/10.13026/C2XW26>.
- Sarvesh Soni and Dina Demner-Fushman. 2026. *ArchEHR-QA: A Dataset for Addressing Patient’s Information Needs related to Clinical Course of Hospitalization*. PhysioNet. PID <https://doi.org/10.1038/s41597-026-06639-z>.
- U.S. National Library of Medicine. 2024. *UMLS – Unified Medical Language System*. U.S. National Library of Medicine. PID <https://www.nlm.nih.gov/research/umls/>.

## A. Training Setup

We train LoRA adapters with TRL’s SFTTrainer and DPOTrainer (Hugging Face, 2024). Unless otherwise noted, runs use bf16, gradient checkpointing, AdamW, and sequence length 512. Full-train runs (for test submissions) use all available labeled pairs without a dev split and omit evaluation during training. Training is executed sequentially and logged per run under `runs/<approach>/launcher.log`.

- SFT: learning rate  $2 \times 10^{-5}$ , 1 epoch, batch size 1, gradient accumulation 4.
- LoRA (Qwen):  $r = 64$ ,  $\alpha = 128$ , dropout 0.05; target modules `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`.
- LoRA (GPT-OSS):  $r = 16$ ,  $\alpha = 64$ , dropout 0.05.
- DPO:  $\beta = 0.1$ , 2 candidates per prompt; full-train uses all rows.
- Distillation (Subtask 1 exploratory): Claude pseudo labels merged with ArchEHR + MeQ-Sum.
- Retrieval: top- $k=6$ , 4 exemplars; embeddings from SentenceTransformers; clinical variant uses BioBERT.
- Hardware: NVIDIA H100 80GB GPUs.

## B. Run Inventory (Subtask 1)

Table 9 lists all Subtask 1 configurations evaluated during development and submitted to the official test phase. Prompting-based runs (zero-shot and few-shot) serve as baselines; retrieval-augmented variants build on these by replacing fixed exemplars with dynamically retrieved nearest neighbors. Supervised fine-tuning and DPO runs use LoRA adapters trained on the released labeled pairs, with the full-train DPO variant submitted to the test set. All runs share the same post-processing pipeline to enforce the 15-word output constraint.

## C. Run Inventory (Subtask 2)

Table 10 documents the three configurations submitted for Subtask 2. All runs use the same hybrid similarity-based example selection strategy and identical prompt template, isolating the effect of model choice and the presence or absence of in-context demonstrations. The ensemble runs combine GPT-OSS-120B and Qwen3-30B-A3B via a union merge; the Claude Opus 4.6 run serves as a single-model proprietary reference under otherwise identical conditions.

## D. Run Inventory (Subtask 3)

Table 11 provides the full run inventory for Subtask 3, covering all 10 configurations evaluated on the development set and the three submitted to the official test phase (marked in bold). The inventory spans four paradigms: zero-shot prompting, fixed few-shot prompting, retrieval-augmented few-shot prompting, and parameter-efficient fine-tuning via QLoRA, with an additional DPO variant. GPT-OSS-120B variants follow the same prompting templates as the corresponding Qwen runs and are omitted from the table for brevity. All runs apply identical 75-word post-processing and use gold clinician questions as input.

## E. Leaderboards

Table 12 places this submission in the context of the official Codabench leaderboard, where our system ranked second out of 16 participating teams with a micro-F1 of 81.3.

| Approach                     | Data                             | LLM             | Training/Inference           | Notes                           |
|------------------------------|----------------------------------|-----------------|------------------------------|---------------------------------|
| prompt_zeroshot              | Released labeled inputs (21–120) | Qwen3-30B-A3B   | temp 0.2, top-p 0.9, 2 seeds | base prompt                     |
| prompt_fewshot               | Released labeled inputs (21–120) | Qwen3-30B-A3B   | temp 0.2, top-p 0.9, 2 seeds | 4 exemplars from training pairs |
| prompt_zeroshot_claude       | Released labeled inputs (21–120) | Claude Opus 4.5 | temp 0.2, 2 seeds            | closed LLM baseline             |
| <b>prompt_fewshot_claude</b> | Released labeled inputs (21–120) | Claude Opus 4.5 | temp 0.2, 2 seeds            | 4 exemplars                     |
| <b>rag_fewshot</b>           | Released labeled inputs (21–120) | Qwen3-30B-A3B   | temp 0.2, top-p 0.9          | FAISS all-mpnet; top-k 6        |
| rag_fewshot_clinical         | Released labeled inputs (21–120) | Qwen3-30B-A3B   | temp 0.2, top-p 0.9          | BioBERT embed + rerank          |
| sft_qwen_instruct            | Train pairs + aug                | Qwen3-30B-A3B   | LoRA r64, lr 2e-5, ep1       | dev split (case-id)             |
| sft_gptoss                   | Train pairs + aug                | gpt-oss-120b    | LoRA r16, lr 2e-5, ep1       | chat template                   |
| sft_qwen_distill             | Distill pairs                    | Qwen3-30B-A3B   | LoRA r64, lr 2e-5, ep1       | pseudo labels                   |
| <b>dpo_qwen_fulltrain</b>    | All pairs                        | Qwen3-30B-A3B   | beta 0.1, 2 candidates       | full-train for test             |

Table 9: Run inventory for Subtask 1. Bold approaches were submitted to the official test phase. All runs use max 64 new tokens and post-processing to enforce  $\leq 15$  words.

| Approach       | Data                                                       | LLM                                  | Inference                                | Notes                                                                        |
|----------------|------------------------------------------------------------|--------------------------------------|------------------------------------------|------------------------------------------------------------------------------|
| opus_5shot     | Dev (20 cases) as few-shot pool; test split for submission | Claude Opus 4.6                      | temp 0.0, greedy decoding, Anthropic API | 5 exemplars via hybrid similarity ( $\alpha=0.7$ ); gold clinician questions |
| ensemble_5shot | Dev (20 cases) as few-shot pool; test split for submission | GPT-OSS-120B + Qwen3-30B-A3B (union) | temp 0.0, greedy decoding, vLLM          | 5 exemplars via hybrid similarity; None-elimination for GPT-OSS; union merge |
| ensemble_0shot | Test split for submission                                  | GPT-OSS-120B + Qwen3-30B-A3B (union) | temp 0.0, greedy decoding, vLLM          | No in-context examples; None-elimination for GPT-OSS; union merge            |

Table 10: Run inventory for Subtask 2. Hybrid similarity combines all-MiniLM-L6-v2 embeddings (weight 0.7) and ROUGE-L (weight 0.3). All runs use gold clinician questions and regex-based output parsing.

| Approach               | Data                 | LLM             | Inference                            | Notes                                        |
|------------------------|----------------------|-----------------|--------------------------------------|----------------------------------------------|
| prompt_zeroshot        | Train (gold answers) | Qwen3-30B-A3B   | temp 0.3, top-p 0.9, 2 seeds, rerank | Zero-shot; 75-word post-processing           |
| prompt_fewshot         | Train (gold answers) | Qwen3-30B-A3B   | temp 0.3, top-p 0.9, 2 seeds, rerank | 4 fixed exemplars (seed 13)                  |
| rag_fewshot            | Train (gold answers) | Qwen3-30B-A3B   | temp 0.3, top-p 0.9, 2 seeds, rerank | FAISS mpnet top-6→4                          |
| prompt_zeroshot_claude | Train (gold answers) | Claude Opus 4.5 | temp 0.3, 1 seed, API                | <b>Submitted to test</b>                     |
| prompt_fewshot_claude  | Train (gold answers) | Claude Opus 4.5 | temp 0.3, 1 seed, API                | 4 fixed exemplars                            |
| rag_fewshot_claude     | Train (gold answers) | Claude Opus 4.5 | temp 0.3, 2 seeds, API               | <b>Submitted to test</b> ; FAISS mpnet       |
| sft_qwen_instruct      | Train (gold answers) | Qwen3-30B-A3B   | QLoRA 4-bit; LoRA $r=8$ , 5 ep       | Best ckpt step 35                            |
| sft_qwen_instruct_long | Train (gold answers) | Qwen3-30B-A3B   | QLoRA 4-bit; LoRA $r=16$ , 10 ep     | <b>Submitted to test</b> ; best ckpt step 70 |
| dpo_qwen               | Train (gold answers) | Qwen3-30B-A3B   | DPO $\beta=0.1$ ; LoRA $r=8$ , 3 ep  | ROUGE-L preference pairs                     |
| dpo_qwen_full          | Train (gold answers) | Qwen3-30B-A3B   | DPO $\beta=0.1$ ; LoRA $r=16$ , 5 ep | Merged adapter at inference                  |

Table 11: Run inventory for Subtask 3. GPT-OSS-120B variants (zero-shot, few-shot, RAG) follow the same template as the corresponding Qwen runs. All runs use gold clinician questions and 75-word post-processing.



| # | Team             | $P_{mi}$    | $R_{mi}$    | $F1_{mi}$   |
|---|------------------|-------------|-------------|-------------|
| 1 | BIT.UA-AAUBS     | <b>88.0</b> | 75.9        | <b>81.5</b> |
| 2 | Ours             | <u>86.9</u> | 76.3        | <u>81.3</u> |
| 3 | Yale-DM-Lab      | 83.3        | <u>77.7</u> | 80.4        |
| 4 | OptiMed          | 80.7        | <b>79.8</b> | 80.3        |
| 5 | UIC-AIHealth4All | 83.6        | 76.3        | 79.8        |

Table 12: Subtask 4 Codabench leaderboard (top-5 of 16 teams). Subscripts  $_{mi}$  denote micro-averaged metrics.

# sebis at ArchEHR-QA 2026: How Much Can You Do Locally? Evaluating Grounded EHR QA on a Single Notebook

Ibrahim Ebrar Yurt<sup>1\*</sup>, Fabian Karl<sup>1\*</sup>, Tejaswi Choppa<sup>1</sup>, Florian Matthes<sup>1</sup>

<sup>1</sup>Technical University of Munich, School of Computation, Information and Technology, Germany  
{ibrahim.yurt, fabian.karl, tejaswi.choppa, matthes}@tum.de

\*Shared first authorship

## Abstract

Clinical question answering over electronic health records (EHRs) can help clinicians and patients access relevant medical information more efficiently. However, many recent approaches rely on large cloud-based models, which are difficult to deploy in clinical environments due to privacy constraints and computational requirements. In this work, we investigate how far grounded EHR question answering can be pushed when restricted to a single notebook. We participate in all four subtasks of the ArchEHR-QA 2026 shared task and evaluate several approaches designed to run on commodity hardware. All experiments are conducted locally without external APIs or cloud infrastructure. Our results show that such systems can achieve competitive performance on the shared task leaderboards. In particular, our submissions perform above average in two subtasks, and we observe that smaller models can approach the performance of much larger systems when properly configured. These findings suggest that privacy-preserving EHR QA systems running fully locally are feasible with current models and commodity hardware.

The source code is available at <https://github.com/ibrahimey/ArchEHR-QA-2026>.

**Keywords:** Electronic Health Records, Clinical Question Answering, Local Language Models

## 1. Introduction

Electronic Health Records (EHRs) contain extensive clinical information about patients, including physician notes, laboratory results, medication histories, and diagnostic reports. Efficient access to this information is critical for both clinicians and patients. Question Answering (QA) over EHRs aims to provide natural language interfaces that retrieve clinically relevant information directly from medical records, thereby supporting clinical decision making and reducing physician workload (Bardhan et al., 2024; Pampari et al., 2018).

In clinical practice, both physicians and patients frequently ask about patient histories, treatments, medications, and diagnostic results (Pampari et al., 2018). Automating responses to such questions can improve efficiency and reduce cognitive load on healthcare professionals. Importantly, QA systems that ground their answers in explicit evidence from medical records can further improve transparency and trust in automated systems, especially in high-stakes domains such as healthcare.

Despite recent advances in large language models (LLMs), applying them to EHR QA presents major practical challenges. Clinical records contain highly sensitive personal health information and are subject to strict privacy regulations such as HIPAA and GDPR. Thus, medical institutions are often unable to send EHR data to external cloud services for processing (Jonnagaddala and Wong, 2025). Meanwhile, modern LLM-based QA systems typically rely on large models that require specialized hardware accelerators or cloud-based inference.

In practice, this creates a gap between research and deployable clinical systems. Many healthcare environments lack the infrastructure that is needed to host large models (Oke et al., 2025). Instead, physicians often have to rely on standard workstation or notebook hardware. For real-world adoption, EHR QA systems must therefore be able to run entirely on local devices while still delivering acceptable performance.

While extensive research has focused on improving language models for biomedical and clinical tasks (Singhal et al., 2023, 2025), relatively little work has investigated how such methods perform under strict local deployment constraints (Bardhan et al., 2024; Blašković et al., 2025). Most state-of-the-art systems rely on large computational resources or cloud-hosted APIs. Consequently, there is a research gap in understanding which architectures and strategies remain effective even when all components must be run on a single local device.

In this work, we investigate how far grounded EHR question answering can be pushed using only locally executable models. We participate in all subtasks of the ArchEHR-QA (Soni and Demner-Fushman, 2026b) shared task and evaluate several strategies designed to operate on commodity hardware. Our experiments explore three classes of approaches: (1) fine-tuned BERT-style classifiers, (2) embedding-based methods, and (3) small or quantized language models for answer generation. All experiments, including training, inference, and evaluation, are performed locally on commodity hardware without relying on external APIs or cloud infrastructure.

Our key contributions are:

- **EHR QA on Commodity Hardware:** We demonstrate that a complete clinical QA pipeline can run entirely on standard commodity hardware.
- **Strong Embedding Baselines:** We show that out-of-the-box dense embedding models surprisingly beat fine-tuned cross-encoders for evidence extraction and alignment.
- **Synthetic Data for Encoder Tuning:** We highlight the brittleness of BERT-style classifiers in low-resource clinical settings and propose a local LLM-based synthetic data pipeline to help stabilize their training.
- **Effectiveness of Quantized LMs:** We establish that quantized and small language models are highly capable of handling the generative question interpretation and answer formulation tasks locally.

## 2. Related Work

### 2.1. EHR Question Answering

EHR question-answering systems have largely been developed and evaluated under the assumption of abundant computational resources, leaving the question of local, privacy-preserving deployment underexplored. Early EHR QA systems were largely framed as reading comprehension and extraction tasks, where systems were asked to identify answer spans in the clinical notes (Bardhan et al., 2024). This formulation was especially prominent in work built on emrQA (Pampari et al., 2018), which established span extraction as a common evaluation setup for clinical question answering and used a DrQA-style (Chen et al., 2017) recurrent document reader as a baseline. Subsequent work moved from the earlier recurrent neural network based document readers towards transformer-based extractive QA models. In particular, Soni and Roberts (2020) evaluated BERT (Devlin et al., 2018), BioBERT (Lee et al., 2019), and ClinicalBERT (Huang et al., 2019) on emrQA in a machine reading comprehension task. They found that intermediate fine-tuning on SQuAD (Rajpurkar et al., 2016) improved answer-span prediction and that ClinicalBERT achieved the strongest emrQA result under sequential fine-tuning. Further, Lanz and Pecina (2024) proposed a two-step retrieve-then-read pipeline where the long clinical records were first segmented into paragraphs, and a retrieval model selected the most relevant segment. A QA model then extracted the answer from the retrieved context. However, Kweon et al. (2024) argued that the realistic clinical QA was more complex than

the earlier extractive benchmarks suggested. They noted that prior datasets largely framed EHR QA as span extraction from a single note, whereas real patient-specific questions might require synthesizing information across multiple clinical notes. To address this, EHRNoteQA (Kweon et al., 2024) introduced patient-specific questions spanning ten topics, including questions that require information from multiple discharge summaries. With the rise of large language models, generative approaches also became more prominent, and they evaluated 27 LLMs in both open-ended and multiple-choice settings. The ArchEHR-QA shared task (Soni and Demner-Fushman, 2026a) pushed this further by requiring answers to be explicitly grounded in clinical evidence, with each answer sentence accompanied by sentence-level citations to the relevant sentences of the clinical note.

Despite this progress, the current LLM-based approaches still remain difficult to deploy in real clinical settings due to the computational costs and privacy concerns.

### 2.2. Local and Privacy Preserving Clinical NLP

Some of the recent work demonstrates that the privacy-preserving local deployment for clinical NLP is feasible in practice. Griot et al. (2025) implemented a GDPR-compliant LLM assistant directly within a live hospital EHR system showing that effective clinical NLP not only depends on model quality but also on security, governance, and workflow integration. Blašković et al. (2025) similarly analyzed the trade-offs between local and hosted LLMs, arguing that on-premise models can improve privacy, latency, and compliance but at the cost of greater computational and operational burden. Although recent work has begun to address privacy-preserving deployment in clinical NLP, there is still limited work on how grounded EHR QA methods compare when all training and inference must remain on commodity hardware.

## 3. Methods

### 3.1. Task Formulation

The shared task focuses on evidence-grounded question answering over electronic health records (EHRs). Given a question  $q$  and a clinical document  $D$  consisting of sentences  $\{s_1, \dots, s_n\}$ , the system must identify textual evidence supporting the answer and generate a concise response grounded in that evidence. The task is split into four subtasks: question interpretation (Subtask 1), evidence identification (Subtask 2), answer generation (Subtask 3), and evidence alignment (Subtask 4).

**Question Interpretation** Patient-authored questions are often verbose and unstructured, intertwining complex personal narratives with medical queries. This subtask requires models to transform the raw patient narrative  $q$  into a concise, clinician-interpreted question  $q_{clin}$ , restricted to 15 words. The objective is to distill the core clinical information needed into a targeted query that a clinician would write.

**Evidence Identification** Clinical notes provide dense, multi-faceted context spanning various events and diagnoses. Given the question ( $q$  or  $q_{clin}$ ) and the segmented clinical document  $D = \{s_1, \dots, s_n\}$ , systems must extract a minimal and sufficient evidence subset  $E \subseteq D$  necessary to formulate an answer.

**Answer Generation** This subtask challenges models to synthesize a coherent, patient-friendly answer  $A$  consisting of generated sentences  $\{a_1, \dots, a_m\}$ , restricted to a maximum length of 75 words. The generated response  $A$  must directly address the query  $q$  while remaining strictly grounded in the clinical document  $D$ .

**Evidence Alignment** The final subtask enforces explicit traceability by aligning the generated text  $A$  back to the source  $D$ . Formulated as a many-to-many mapping problem, models must link each answer sentence  $a_i \in A$  to a specific set of supporting evidence sentences (citations)  $C_{a_i} \subseteq D$ .

### 3.2. Architecture Design

**Question Interpretation** We use three models for this subtask: Qwen3-4B (Qwen Team, 2025), Qwen2.5-14B (Qwen Team, 2024), and gpt-oss-120b (OpenAI, 2025).

We evaluate different few-shot prompting setups using the first three or five cases from the development set. In particular, we test two 3-shot prompts for Qwen3-4B, three 3-shot prompts for Qwen2.5-14B, and for gpt-oss-120b we consider three 3-shot prompts and four 5-shot prompts.

Inspecting the outputs of Qwen2.5-14B shows that some answers far exceed the 15 word limit. Hence, we conduct experiments with a two-step approach where the initial answer is provided by a 3-shot prompt using Qwen2.5-14B, and is revised by Qwen3-4B to obtain the final answer.

Finally, following Leviathan et al. (2025), we also experiment with query repetition using gpt-oss-120b with the best-performing prompt on the development set.

**Evidence Identification** To identify sentences in the clinical note relevant to answering a question,

we explore both retrieval-based approaches and supervised classification methods.

To address the limited size of the development set for fine-tuning encoder models, we generate a supplementary synthetic dataset. Using a local deployment of Llama3.1-70B (Grattafiori et al., 2024), we synthesize 10 novel cases for each original development case, resulting in 200 new synthetic cases comprising 1,818 sentences annotated with three-class relevance labels. This generation process utilizes a two-stage pipeline consisting of an initial synthesis followed by targeted LLM-based repairs, strictly guided by manually defined quality thresholds (e.g., bounding sentence length to 10-500 characters and restricting the ratios of essential and supplementary relevance labels). Details are reported in Appendix B.

Utilizing this synthetic dataset, we train a cross-encoder evidence classifier by fine-tuning a BERT-style encoder. To capture nuanced relevance, the architecture features a shared representation layer that feeds into two distinct task heads at once: a 3-way fine-grained classification head and a 2-way binary classification head, following the multi-head training approach of HYDRA (Karl and Scherp, 2025). To prevent the synthetic cases from dominating the training signal, we balance the dataset by upsampling the real cases and downsampling the synthetic data to achieve a strict 1:1 ratio. The optimal configuration is robustly validated using case-level K-fold cross-validation.

Regarding retrieval-based approaches, we use two models with different architectures and parameter counts, namely Qwen3-Embedding-8B (Zhang et al., 2025), and MedCPT-Cross-Encoder (Jin et al., 2023), to obtain similarity scores between  $q_{clin}$  and each  $s_i \in D$ . Then, we determine decision thresholds  $t$  based on strict micro F1 scores on the development set for different  $t$  values, defining the set of sentences  $E = \{s_i \in D \mid \text{score}(q_{clin}, s_i) > t\}$  as relevant. See Figure 1 for the relationship between different threshold values and F1 score.

MedCPT-Cross-Encoder provides pair-wise scores for a query ( $q_{clin}$ ) and each  $s_i \in D$ , and these scores are used directly. For Qwen3-Embedding-8B embedding model, the scores are computed as the cosine similarity between the embeddings of  $q_{clin}$  and each  $s_i \in D$ .

**Answer Generation** We experiment with two setups for this subtask: a 0-shot prompt that includes the patient-authored question  $q$ , the clinician-interpreted question  $q_{clin}$ , and the clinical note excerpt  $D$  with gpt-oss-120b; and a two-step 0-shot approach using  $q$  and  $D$  with Qwen3-4B.

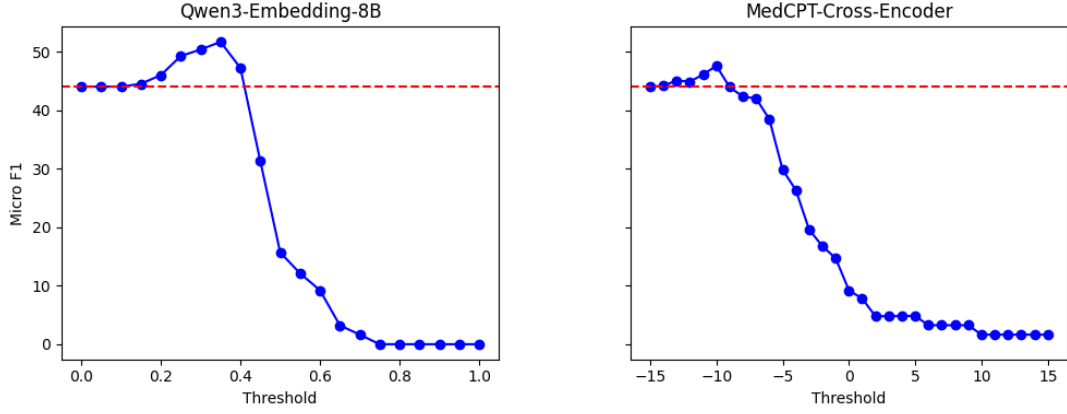


Figure 1: Overall scores for different threshold values for both methods on the development set for Subtask 2. Dashed red line shows the naive baseline of classifying everything as relevant.

**Evidence Alignment** To identify the supporting evidence for each answer sentence, we explore both out-of-the-box approaches using embedding and generative models, as well as supervised classification methods.

For fine-tuning our alignment cross-encoder, we use a BERT-style architecture to model the pairwise relationships among the queries ( $q$  and  $q_{clin}$ ), generated answer  $A$ , and source evidence  $D$ . Specifically, this classification is performed between each  $a_i \in A$  and  $s_j \in D$ . In contrast to the multi-head setup used in Subtask 2, we employ a standard binary classification head to output a relevance decision, indicating whether the specific evidence sentence supports the answer sentence, since the task does not allow for a finer-grained distinction. In addition, we experiment with expanding the training corpus by reusing the synthetic data generated for evidence identification (Subtask 2) as pseudo-alignments, maintaining the same 1:1 ratio between real and synthetic data to preserve training balance.

For the out-of-the-box models, we consider four approaches. First, we adopt a threshold approach similar to Subtask 2 using Qwen3-Embedding-8B, computing similarities between each  $s_i \in D$  and  $a_j \in A$ . Second, we apply two-step prompting for list-wise alignment using Qwen3-4B, where  $q$ ,  $q_{clin}$ ,  $D$ , and  $A$  are provided in the prompt. The first step uses a 1-shot prompt to generate a natural language response, and the second step reformats the output into JSON. Third, we perform pair-wise alignment using a 0-shot prompt as a binary classifier with Qwen3.5-35B-A3B (Qwen Team, 2026). For each  $a_i \in A$  and  $s_j \in D$ , the model determines whether  $s_j$  supports  $a_i$ . Fourth, we use list-wise alignment with a 1-shot prompt and Qwen3.5-35B-A3B, similar to the second approach but without the reformatting step.

For the 1-shot prompts, the first case from the

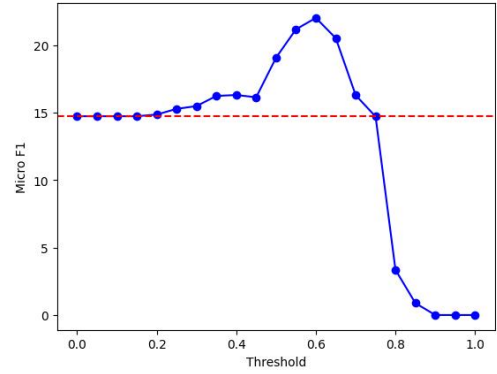


Figure 2: Overall scores for different threshold values on the development set for Subtask 4. Dashed red line shows the naive baseline of classifying everything as relevant.

development set is used as the example. See Figure 2 for the relationship between threshold values and F1 for the Qwen3-Embedding-8B approach.

## 4. Experimental Apparatus

### 4.1. Dataset

We conduct all experiments using the ArchEHR-QA 2026 dataset (Soni and Demner-Fushman, 2026a), designed to benchmark question answering over electronic health records (EHRs). The dataset models an interaction scenario in which patients ask questions about their health records and clinicians provide answers with explicit evidence grounding.

Each instance in the dataset, referred to as a *case*, combines a patient question with a clinical note excerpt from the MIMIC database (Johnson et al., 2016). Alongside the raw text, the dataset provides multiple layers of expert annotations supporting different stages of the QA pipeline.



Every case includes a free-form *patient question*, a shorter *clinician-interpreted question* (reference for Subtask 1), a *clinical note excerpt* segmented into numbered sentences to enable sentence-level grounding (reference for Subtask 2), a *clinician-written answer* addressing the question (reference for Subtask 3), *evidence links* connecting answer sentences to supporting sentences from the clinical note excerpt (reference for Subtask 4), and the *clinical specialty* associated with the case.

The dataset is divided into a development set with 20 cases and two test sets, *test* and *test-2026*, comprising 100 and 47 cases, respectively. The labels for the test sets are withheld, and a maximum of three submissions can be evaluated. The *test-2026* set is used for Subtasks 1, 2, and 3, while Subtask 4 is evaluated on both test sets. Dataset statistics are summarized in Table 1.

## 4.2. Models

To address the diverse requirements of the four ArchEHR-QA subtasks while adhering to our local hardware constraints, we evaluate a broad spectrum of parameter-efficient model architectures. An overview of all utilized models and their respective parameter counts is provided in Table 2.

**Autoregressive Generative Models** For tasks requiring text generation, we experiment with several large and small instruction-tuned autoregressive language models (i.e., Qwen3-4B-Instruct (Qwen Team, 2025), Qwen2.5-14B-Instruct (Qwen Team, 2024), Qwen3.5-35B (Qwen Team, 2026), and gpt-oss-120b (OpenAI, 2025)). To enable local execution of larger models, we run quantized versions converted to the MLX framework (Hannun et al., 2023) with 4-bit precision.

**Decoder-only Embedding Models** For similarity-based retrieval and alignment tasks, we utilize the Qwen3-Embedding-8B (Zhang et al., 2025) model. This model produces dense vector representations, enabling efficient semantic similarity computation using cosine similarity. Additionally, we employ the MedCPT cross-encoder (Jin et al., 2023), a domain-specific model trained on biomedical literature, to compute direct pairwise relevance scores between queries and clinical text.

**Encoder-only Classification Models** For binary and multi-class sentence classification, we rely on parameter-efficient, BERT-style encoder architectures. These models are well-suited for local fine-tuning and rapid cross-encoder inference and are known to be strong short-text classifiers (Karl and Scherp, 2023). We evaluate domain-specific models and strong general-domain encoders.

## 4.3. Hyperparameter Optimization

For all generative inference-only experiments, we use the default decoding parameters, with a temperature of 0.7 and a top-p value of 0.9. No additional tuning is performed for these settings.

For the fine-tuning of our classifier in Subtask 2: Evidence Identification and Subtask 4: Evidence Alignment, we conduct a grid search to identify optimal hyperparameters. The hyperparameter search space includes learning rates in  $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$ , training epochs in  $\{1, 2\}$ , and dropout probabilities in  $\{0.10, 0.20\}$ . The optimal parameters are determined by evaluating performance across 5-fold cross-validation runs on the development set. The selected hyperparameters are reported in Table 3.

## 4.4. Metrics

To evaluate system performance across the four subtasks, we employ a combination of established classification and text-generation metrics.

For the generative tasks (Subtasks 1 and 3), we report several evaluation metrics to capture various aspects of text quality. We use, BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) SARI (Xu et al., 2016), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023). Beyond lexical overlap, we employ BERTScore (Zhang et al., 2020) to compute the semantic similarity between the generated and reference sentences using contextual embeddings.

For the extraction and alignment tasks (Subtasks 2 and 4), performance is measured using standard micro Precision, micro Recall, and micro F1 scores.

We strictly followed the shared task’s official evaluation scripts for all subtasks. For Subtask 4, note that the script computes metrics in a specific manner that may appear counterintuitive (e.g., not reaching 100% recall in the all-relevant configurations), as defined by the task organizers.

## 4.5. Hardware

A core motivation of our work is to demonstrate the feasibility of executing EHR question-answering pipelines entirely on commodity hardware. All experiments are therefore conducted locally on Apple Silicon devices rather than GPU clusters.

Most of our experiments are performed on an Apple MacBook M4 Pro with 48GB of unified memory. This system is sufficient for training the classifier models and running most inference experiments.

To further evaluate the feasibility of running larger models locally, we conduct experiments on high-end consumer hardware. Specifically, we use an Apple Mac Studio M3 Max with 96GB of memory to run experiments with the gpt-oss-120B model.

| Statistic                         | Development | Test   | Test-2026 |
|-----------------------------------|-------------|--------|-----------|
| Cases                             | 20          | 100    | 47        |
| Note sentences                    | 428         | 2599   | 1596      |
| Average note sentences per case   | 21.40       | 25.99  | 33.96     |
| Average note sentence length      | 15.00       | 14.64  | 14.64     |
| Average note length per case      | 320.90      | 380.42 | 497.26    |
| Average patient question length   | 16.20       | 14.64  | 14.72     |
| Average clinician question length | 10.75       | 10.63  | 10.43     |
| Average answer sentences per case | 4.75        | 4.70   | 4.40      |
| Average answer sentence length    | 15.49       | 15.40  | 16.51     |

Table 1: Descriptive statistics of the ArchEHR-QA dataset across the development and two test splits. All text length measurements are reported as word counts.

| Model                                        | #Parameter               |
|----------------------------------------------|--------------------------|
| <i>Autoregressive Generative Models</i>      |                          |
| mlx-community/gpt-oss-120b-MXFP4-Q4          | 117B total / 5.1B active |
| RepublicOfKorokke/Qwen3.5-35B-A3B-mlx-lm-mx4 | 35B total / 3B active    |
| Qwen/Qwen2.5-14B-Instruct                    | 14.7B                    |
| Qwen/Qwen3-4B-Instruct-2507                  | 4B                       |
| <i>Retrieval Models</i>                      |                          |
| Qwen/Qwen3-Embedding-8B                      | 8B                       |
| ncbi/MedCPT-Cross-Encoder                    | 109M                     |
| <i>Classification Models</i>                 |                          |
| microsoft/deberta-base                       | 138M                     |
| emilyalsentzer/Bio_ClinicalBERT              | 108M                     |

Table 2: Evaluated models categorized by architecture type and sorted by decreasing parameter count. All models are obtained from Hugging Face (B: billions, M: millions).

## 5. Results

**Question Interpretation** We select the three best-performing setups on the development set, based on BERTScore, for evaluation on the test set. Namely, gpt-oss-120b with a 5-shot prompt, the repeated-query approach with the same prompt, and the two-step approach using smaller models. Table 4 shows the development and test set results for these models. Overall score corresponds to the average of the four scores. See Appendix A for the prompt template corresponding to the single-query setup, which achieved the best overall score.

**Evidence Identification** For this subtask, we compare our methods to the naive baseline of considering  $E = D$ , that is considering every sentence as relevant evidence on the development set. Surprisingly, most methods struggle to outperform this baseline substantially. See Table 5 for strict micro precision, recall, and F1 scores for every method on the development set. Based on these results, we selected Qwen3-Embedding-8B, and Bio\_ClinicalBERT-synth-HYDRA approaches to be evaluated on the test set, where they had overall scores of 51.61 and 44.43 respectively, which can also be seen on the same table.

**Answer Generation** Table 6 shows development results for both approaches and test results for the approach with the best performance (gpt-oss-120b). The prompt template is included in Appendix A.

**Evidence Alignment** Similar to Evidence Identification, we again compare our methods to the naive baseline of considering  $C_{a_i} = D, \forall a_i \in A$ ; that is considering every excerpt sentence as supporting evidence for every answer sentence, on the development set. See Table 7 for strict micro precision, recall, and F1 scores for every method on the development set. Based on these results, we selected Qwen3-Embedding-8B, Qwen3.5-35B list-wise, and Bio\_ClinicalBERT approaches to be evaluated on the test set, where they had overall scores of 59.45, 74.84, and 8.33 respectively, which can also be seen on the same table. See Appendix A for the prompt template for the setup with the best overall score (Qwen3.5-35B list-wise).

## 6. Discussion

Results from Subtasks 1 and 3 indicate that although larger models generally achieve stronger performance, the performance gap compared to smaller models remains moderate. Comparable re-

| Task      | Model            | LR                 | Epochs | Dropout |
|-----------|------------------|--------------------|--------|---------|
| Subtask 2 | Bio_ClinicalBERT | $2 \times 10^{-5}$ | 1      | 0.1     |
| Subtask 2 | DeBERTa-base     | $2 \times 10^{-5}$ | 1      | 0.1     |
| Subtask 4 | Bio_ClinicalBERT | $3 \times 10^{-5}$ | 1      | 0.1     |
| Subtask 4 | DeBERTa-base     | $2 \times 10^{-5}$ | 1      | 0.1     |

Table 3: Optimized hyperparameters for the BERT-style classifiers in Subtask 2 and Subtask 4. Values are selected via a grid search evaluated with 5-fold cross-validation on the development set.

| Metric          | SQ           | DQ           | TS           |
|-----------------|--------------|--------------|--------------|
| <i>Dev Set</i>  |              |              |              |
| ROUGE-Lsum      | 23.51        | 24.24        | <b>28.74</b> |
| BERTScore       | 41.15        | <b>43.79</b> | 41.94        |
| <i>Test Set</i> |              |              |              |
| ROUGE-Lsum      | 22.92        | <b>23.09</b> | 21.19        |
| BERTScore       | 36.93        | <b>37.29</b> | 33.10        |
| AlignScore      | 21.42        | 20.22        | <b>22.91</b> |
| MEDCON (UMLS)   | 21.26        | <b>21.82</b> | 21.18        |
| Overall Score   | <b>25.63</b> | 25.61        | 24.59        |

Table 4: Test and development set results for Subtask 1. Only ROUGE and BERTScore are used on the development set. SQ: Single Query, DQ: Double Query, TS: Two Step.

sults can therefore be obtained with much smaller architectures, suggesting that model scale alone is not the main driver of performance for these tasks.

For Subtask 1, we evaluated prompt repetition as proposed by Leviathan et al. (2025). In contrast to their findings, in our case the prompt repetition did not considerably outperform the single-query approach. Nevertheless, it produced slightly higher results on three out of the four evaluation metrics on the test set, indicating a modest but consistent improvement across most metrics.

The results for Subtask 2 reveal a strong recall bias across many approaches. Notably, the baseline that classifies every candidate as relevant achieves perfect recall and surprisingly competitive F1 performance. This observation suggests that the dataset implicitly rewards high recall and that reliably identifying truly irrelevant contexts remains challenging. More advanced methods, such as embedding similarity or cross-encoder ranking, generally increase precision but often at the cost of reduced recall. Interestingly, models trained with HYDRA and synthetic data exhibit somewhat more balanced precision–recall trade-offs in certain cases. This may indicate that these models benefit from additional training signals and that synthetic examples can help establish clearer decision boundaries when real training data is limited.

A comparison between DeBERTa and Bio\_ClinicalBERT further highlights the importance of domain-specific modeling. The

biomedical model consistently performs better in Subtask 2 than the general-purpose model, suggesting that the task contains considerable domain-specific terminology and contextual cues that are better captured by specialized pretraining.

Unlike the other subtasks, Subtask 4 exhibits a substantial discrepancy between the development and test set performance. We hypothesize that this may be caused by a distribution shift between the two splits. This could also explain the severe degradation observed for the fine-tuned model on the test set compared to out-of-the-box models. Since the fine-tuned models were trained on the full development set using hyperparameters derived from k-fold validation, they may have overfitted to the development distribution, which does not fully reflect the characteristics of the test set.

Similar to Subtask 2, the biomedical model again outperforms the general-domain model in Subtask 4. However, unlike Subtask 2, synthetic data improves performance for the general-domain model (DeBERTa) but not for the biomedical model (Bio\_ClinicalBERT). Considering the more balanced precision–recall behavior observed in Subtask 4, this suggests that synthetic data may be particularly beneficial when model predictions are highly skewed, as is the case in Subtask 2.

## 6.1. Future Work and Impact

Our findings establish a solid baseline for local EHR QA. However, our setups still rely on high-end commodity hardware. Future work should explore the feasibility of even smaller, highly optimized language models capable of executing on highly constrained edge devices, such as mobile phones or clinical tablets, to further lower the barrier to entry.

Furthermore, this shared task operates under low-resource data, which likely contributes to the generalization challenges observed with our fine-tuned cross-encoders. To mitigate this data scarcity, future research should explore integrating large-scale proxy datasets from related clinical domains alongside more advanced synthetic data generation methods to better stabilize model fine-tuning and improve robustness to distribution shifts.

| Setup                                  | Dev Set  |          |              | Test Set |          |              |
|----------------------------------------|----------|----------|--------------|----------|----------|--------------|
|                                        | Strict P | Strict R | Strict F1    | Strict P | Strict R | Strict F1    |
| <i>Baseline</i>                        |          |          |              |          |          |              |
| Everything is relevant                 | 28.27    | 100.00   | <u>44.08</u> | –        | –        | –            |
| <i>Similarity Threshold</i>            |          |          |              |          |          |              |
| Qwen3-Embedding-8B                     | 48.55    | 55.37    | <b>51.74</b> | 37.92    | 80.74    | <b>51.61</b> |
| MedCPT-Cross-Encoder                   | 34.04    | 79.34    | 47.64        | –        | –        | –            |
| <i>5-Fold Training</i>                 |          |          |              |          |          |              |
| Bio_ClinicalBERT-real-only-binary-head | 31.39    | 93.39    | 46.99        | –        | –        | –            |
| Bio_ClinicalBERT-real-HYDRA            | 32.46    | 92.56    | 48.07        | –        | –        | –            |
| Bio_ClinicalBERT-synth-HYDRA           | 36.01    | 85.12    | <u>50.61</u> | 32.40    | 70.68    | 44.43        |
| DeBERTa-real-only-binary-head          | 28.72    | 94.21    | 44.02        | –        | –        | –            |
| DeBERTa-real-HYDRA                     | 28.27    | 100.00   | 44.08        | –        | –        | –            |
| DeBERTa-synth-HYDRA                    | 28.97    | 93.39    | 44.23        | –        | –        | –            |

Table 5: Subtask 2 results. P and R stand for Precision and Recall, respectively. “Synth” indicates training with synthetic data, while “HYDRA” denotes multiple heads. Development set results for fine-tuned models are computed from out-of-fold predictions. **Bold** values mark the highest overall F1 score, and underlined values show the best F1 within each method category on the development set.

| Metric        | Dev Set |       | Test Set |
|---------------|---------|-------|----------|
|               | Qwen    | gpt   | gpt      |
| Overall Score | 27.76   | 32.38 | 31.51    |
| BLEU          | 1.19    | 4.34  | 3.82     |
| ROUGE-Lsum    | 15.73   | 21.34 | 22.04    |
| SARI          | 49.75   | 58.28 | 56.55    |
| BERTScore     | 30.95   | 39.66 | 39.30    |
| AlignScore    | 41.16   | 38.29 | 33.92    |
| MEDCON (UMLS) | –       | –     | 33.44    |

Table 6: Evaluation results for Subtask 3 on the development and test sets. For the development set MEDCON score was not calculated and therefore was also not included in the overall score, which is the mean of all other scores.

## 7. Conclusion

As shown on the competition leaderboards, our approaches implemented using only commodity hardware, achieve competitive results across multiple subtasks. For Subtask 1<sup>1</sup> and Subtask 4<sup>2</sup>, our submissions perform above the average score. Although our results for Subtask 3<sup>3</sup> are lower, they remain comparable to those of other systems. For Subtask 2<sup>4</sup>, our approach performs less favorably, indicating substantial room for improvement.

Our experiments further demonstrate that, with appropriate configurations and training strategies, smaller models can achieve performance comparable to that of significantly larger models. This high-

lights the potential of resource-efficient approaches for tasks in the clinical domain.

Overall, our findings suggest that privacy-preserving, fully local electronic health record (EHR) question-answering systems are technically feasible using current models and accessible hardware. While such systems may still lag behind large-scale cloud-based solutions in certain aspects, the performance gap appears small enough for many real-world applications in which data privacy and on-device processing are critical requirements.

## 8. Limitations

While our approach demonstrates the feasibility of local EHR QA, several limitations remain. Although our setup avoids external cloud infrastructure, it still relies on high-end commodity hardware (e.g., an Apple MacBook with an M4 Pro chip and 48 GB memory), which may not be readily available in all under-resourced clinical environments. Furthermore, our evaluation is constrained to the ArchEHR-QA dataset, limiting the immediate generalizability of our findings across diverse clinical specialties or multilingual patient populations.

## 9. Ethical Considerations

Our work is primarily motivated by the ethical imperative to protect patient privacy. By executing all models entirely on a local notebook, we eliminate the need to transmit sensitive Protected Health Information to external, third-party cloud APIs. However, the use of generative AI in healthcare carries inherent risks for patient safety. The models we evaluate are prone to hallucinations and false con-

<sup>1</sup>[www.codabench.org/competitions/12865/](http://www.codabench.org/competitions/12865/)

<sup>2</sup>[www.codabench.org/competitions/13528/](http://www.codabench.org/competitions/13528/)

<sup>3</sup>[www.codabench.org/competitions/13527/](http://www.codabench.org/competitions/13527/)

<sup>4</sup>[www.codabench.org/competitions/13526/](http://www.codabench.org/competitions/13526/)



| Model                       | Dev Set |         |              | Test Set |         |              |
|-----------------------------|---------|---------|--------------|----------|---------|--------------|
|                             | Micro-P | Micro-R | Micro-F1     | Micro-P  | Micro-R | Micro-F1     |
| <i>Baseline</i>             |         |         |              |          |         |              |
| Everything is relevant      | 8.16    | 75.11   | <u>14.73</u> | –        | –       | –            |
| <i>Similarity Threshold</i> |         |         |              |          |         |              |
| Qwen3-Embedding-8B          | 22.82   | 21.27   | <u>22.01</u> | 52.41    | 68.69   | 59.45        |
| <i>Prompting</i>            |         |         |              |          |         |              |
| Qwen3-4B                    | 20.00   | 11.76   | 14.81        | –        | –       | –            |
| Qwen3.5-35B pair-wise       | 35.35   | 15.84   | 21.88        | –        | –       | –            |
| Qwen3.5-35B list-wise       | 31.58   | 21.72   | <b>25.74</b> | 79.65    | 70.57   | <b>74.84</b> |
| <i>5-Fold Training</i>      |         |         |              |          |         |              |
| Bio_ClinicalBERT            | 19.14   | 24.43   | <u>21.47</u> | 6.26     | 12.46   | 8.33         |
| Bio_ClinicalBERT + Synth    | 16.27   | 25.33   | 19.82        | –        | –       | –            |
| DeBERTa-base                | 9.24    | 50.67   | 15.63        | –        | –       | –            |
| DeBERTa-base + Synth        | 12.67   | 45.25   | 19.80        | –        | –       | –            |

Table 7: Subtask 4 results. Micro-Precision (P), Micro-Recall (R), and Micro-F1 are reported. “Synth” denotes training augmented with synthetic data. Development set results for fine-tuned models are derived from merged out-of-fold predictions. **Bold** values denote the highest overall F1 score, and underlined values indicate the best F1 score within each method category on the development set.

clusions. In a clinical context, such errors could negatively impact patient care if accepted without careful review.

## 10. Bibliographical References

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:e53636.

Luka Blašković, Nikola Tanković, Ivan Lorencin, and Sandi Baressi Šegota. 2025. Robust clinical querying with local llms: Lexical challenges in nl2sql and retrieval-augmented qa on ehrrs. *Big Data and Cognitive Computing*, 9(10):256.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. [Reading Wikipedia to answer open-domain questions](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

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.

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*.

Maxime Griot, Jean Vanderdonckt, and Demet Yukse. 2025. [Implementation of large language models in electronic health records](#). *PLOS Digital Health*, 4(12):1–18.

Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. 2023. [MLX: Efficient and flexible machine learning on apple silicon](#).

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. [Clinicalbert: Modeling clinical notes and predicting hospital readmission](#). *CoRR*, abs/1904.05342.

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.

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.

Jitendra Jonnagaddala and Zoie Shui-Yee Wong. 2025. Privacy preserving strategies for electronic health records in the era of large language models. *npj Digital Medicine*, 8(1):34.

Fabian Karl and Ansgar Scherp. 2023. [Transformers are short-text classifiers](#). In *Machine*



- Learning and Knowledge Extraction - 7th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2023, Benevento, Italy, August 29 - September 1, 2023, Proceedings*, volume 14065 of *Lecture Notes in Computer Science*, pages 103–122. Springer.
- Fabian Karl and Ansgar Scherp. 2025. [HYDRA: A multi-head encoder-only architecture for hierarchical text classification](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 9292–9303. Association for Computational Linguistics.
- Sunjun Kweon, Jiyou Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwang Hyun Kim, Seunghyun Won, and Edward Choi. 2024. [EHRNoteQA: A Patient-Specific Question Answering Benchmark for Evaluating Large Language Models in Clinical Settings](#). *PhysioNet*. Version 1.0.0.
- Vojtech Lanz and Pavel Pecina. 2024. [Paragraph retrieval for enhanced question answering in clinical documents](#). In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 580–590, Bangkok, Thailand. 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](#). *CoRR*, abs/1901.08746.
- Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2025. Prompt repetition improves non-reasoning llms. *arXiv preprint arXiv:2512.14982*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Femi Oke, Oyeneke Bolaji, and Michael Umakor. 2025. Building ai-ready infrastructure for us healthcare: a product management perspective. *World J Adv Res Rev*, 27(2):588–603.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#).
- Anusri Pampari, Preethi Raghavan, Jennifer J. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2357–2368. 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.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Qwen Team. 2025. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- Qwen Team. 2026. [Qwen3.5: Towards native multimodal agents](#).
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [Squad: 100, 000+ questions for machine comprehension of text](#). *CoRR*, abs/1606.05250.
- 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.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950.
- Sarvesh Soni and Dina Demner-Fushman. 2026a. [A dataset for addressing patient's information needs related to clinical course of hospitalization](#). *Scientific Data*.
- 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.
- Sarvesh Soni and Kirk Roberts. 2020. [Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5532–5538, Marseille, France. European Language Resources Association.
- 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.

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.

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*.

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. *arXiv preprint arXiv:2506.05176*.

## A. Prompt Templates

### Listing 1: Best-performing prompt template for Sub-task 1: Question Interpretation

You rewrite long, messy questions into ONE short core question, phrased in third person from a clinician’s perspective.

Task:

- Identify the single core question being asked.
- Rewrite it as one clear question in THIRD PERSON.

Hard Rules (must follow):

- Output exactly ONE sentence
- Maximum 15 words
- Use third-person only (e.g., the patient, he, she, they)
- Do NOT explain, summarize, or add context
- Do NOT add new information
- Output ONLY the rewritten question

Question Style:

- Use simple forms like ``What'', ``Why'', ``Did'', ``Is'', ``How''

Examples:

<5 Examples>

### Listing 2: Best-performing prompt template for Sub-task 3: Answer Generation

You are a clinical documentation assistant.

Your task is to generate a professional, natural-language answer to the patient’s question using ONLY the information explicitly stated in the provided clinical note excerpt.

Instructions:

- Use only facts supported by the clinical note excerpt.
- Do NOT add outside medical knowledge.
- Do NOT speculate or infer beyond what is documented.
- If the note does not fully answer the question, provide a faithful response limited to the documented information.
- Write in a professional clinical tone.
- Limit your response to a maximum of 75 words (approximately 5 sentences).
- Do not include citations or sentence numbers.
- Do not mention the instructions in your response.

### Listing 3: Best-performing prompt template for Sub-task 4: Evidence Alignment

You are a clinical evidence alignment system.

Your task is to align each answer sentence to supporting clinical note sentences.

Important rules:

1. Alignment is at the answer sentence level.
2. Only include note sentences that DIRECTLY support the answer sentence.
3. Do NOT over-cite.
4. Do NOT under-cite.
5. Do NOT infer beyond the note text.
6. Use only the provided numbered note sentences.
7. If no supporting evidence exists, return an empty list [].
8. Evidence IDs must be strings.
9. Output must be STRICT JSON.
10. Do not include explanations or reasoning.

Output format:

```
[
 {
 "case_id": "<case_id>",
 "prediction": [
 {
 "answer_id": "<ID>",
 "evidence_id": ["<note_id>", ...]
 }
]
 }
]
```

Example:

<1 Example>

## B. Synthetic Data Generation Details

To augment the limited development set for fine-tuning our encoder models, we generate synthetic clinical cases using a local deployment of Llama-70B via Ollama. For each of the 20 original development cases, we generate 10 novel synthetic variations, yielding a total of 200 new cases. This process produces 1,818 synthetic sentences, each assigned a 3-class relevance label ("essential", "supplementary", or "not-relevant").

The generation is executed in a two-stage pipeline: an initial LLM generation phase followed by a targeted LLM-based repair phase to correct formatting and distribution errors. Both steps are strictly guided by the following manually defined quality thresholds to ensure the synthetic data closely mirrors the distribution of the real dataset:

- **Sentence Count constraints:** Between 10 and 20 sentences per case.
- **Sentence Length constraints:** Between 10 and 500 characters per sentence.
- **Relevance Label Ratios:**
  - Essential ratio:  $0.10 \leq r_{ess} \leq 0.40$
  - Supplementary ratio:  $r_{sup} \leq 0.15$
  - Not-Relevant ratio:  $r_{nr} \geq 0.45$

The following prompt template is used during the initial generation phase to construct the synthetic cases based on the provided seed examples from the development set.

Listing 4: Prompt template for generating synthetic clinical QA cases.

You are a medical data  
augmentation assistant. Your

task is to create a synthetic  
clinical note excerpt with  
relevance labels for training  
a medical evidence  
identification model.

TASK: Study the real examples  
below, then generate a NEW  
synthetic case with DIFFERENT  
medical content (different  
condition, treatment, or  
scenario). Pay close attention  
to the label distributions in  
the examples - most sentences  
are "not-relevant".

{examples\_block}

INSTRUCTIONS:

1. Create a NEW patient question about a DIFFERENT medical topic from any of the examples above.
2. Create a corresponding clinician question (concise, max 15 words).
3. Write a LONG clinical note excerpt with 15-30 sentences. Real clinical notes are long and contain many sections (History of Present Illness, Hospital Course, Medications, Labs, Vitals, Plan, etc.). Include realistic clinical detail.
4. Label each sentence as "essential", "supplementary", or "not-relevant":
  - essential: Directly answers the patient's question (~20-35% of sentences)
  - supplementary: Provides helpful context but not critical (~5-10% of sentences, this label is RARE)
  - not-relevant: Does NOT help answer the question (~55-70% of sentences - this should be the MAJORITY)

IMPORTANT DISTRIBUTION GUIDANCE:

- Most sentences in a clinical note are NOT relevant to the patient's specific question. The note covers many topics (vitals, labs, medications, procedures, consults, discharge planning) and only a fraction directly address the question.
- "supplementary" is the RAREST label - only a few sentences

- at most.
- "not-relevant" should be the MOST COMMON label - typically 60%+ of sentences.
- Look at the examples above: notice how the majority of sentences are labeled "not-relevant".

OUTPUT FORMAT (use exactly this JSON format):

```
{
 "patient_question": "...",
 "clinician_question": "...",
 "sentences": [
 { "id": "1", "text": "..." },
 { "id": "2", "text": "..." }
],
 "relevance_labels": [
 { "sentence_id": "1", "relevance": "essential" },
 { "sentence_id": "2", "relevance": "not-relevant" }
]
}
```

Generate a synthetic case now.  
Output ONLY valid JSON, no explanation.

# MedEvi-NS at ArchEHR-QA 2026: Using Clinical Reasoning Principles to Improve Zero-shot Capabilities of Large Language Models in Evidence Alignment

Mengxuan Sun<sup>1\*</sup>, Nicolay Rusnachenko<sup>2\*</sup>

<sup>1</sup> Institute of Applied Health Sciences, University of Aberdeen, UK,

<sup>2</sup> Centre for Applied Creative Technologies (CFACT+), Bournemouth University, UK

m.sun.22@abdn.ac.uk, nrusnachenko@bournemouth.ac.uk

\* These authors contributed equally to this work.

## Abstract

The ArchEHR-QA shared task focuses on grounded question answering using patient EHR data. For the given clinical interpretation of the patient question, note excerpt ( $E$ ) and answer text ( $A$ ), subtask 4 (evidence alignment) aims to cite supporting sentences from  $E$  for each sentence in  $A$ . In this paper, we propose a prompt-engineering methodology that features clinical-reasoning principles in related alignment. We adopt this methodology for GPT-5.2 in zero-shot learning mode. According to our experiments on ArchEHR-QA, incorporating clinical reasoning principles into the prompt improves  $F1_{overall}$  by +2.0%. Our final submission resulted in 77.4% by  $F1_{overall}$ , which positions us at 10<sup>th</sup> out of 16 teams. Our code is publicly available: <https://github.com/nicolay-r/ArchEHR-QA-2026-Task-4-MedEvi-NS>

**Keywords:** Large Language Models, Evidence Alignment, Prompt-based Engineering, Clinical Reasoning

## 1. Introduction

Recent advances in large language models (LLMs) have significantly improved question answering systems across many domains (Shailendra et al., 2024). However, applying such systems to clinical use cases remains challenging as hallucinated or unsupported responses may pose patient safety risks (Kim et al., 2025). In clinical practice, evidence-based medicine emphasises that medical conclusions should be supported by explicit clinical evidence (Lohr et al., 1998). Therefore, ensuring accurate alignment of evidence is essential for maintaining the faithfulness in clinical question answering.

The ArchEHR-QA shared task focuses on grounded question answering using patient EHR data (Soni and Demner-Fushman, 2026a,b). The objective is to answer patient questions and provide evidence supported by clinical notes. In this work, we participate in subtask 4, which requires aligning each answer sentence with the specific supporting sentence(s) in the clinical note excerpt.

Existing approaches to evidence alignment are often formulated as evidence retrieval or semantic matching tasks (Zhao et al., 2024; Karthik et al., 2025; Gupta et al., 2024). However, such approaches may be insufficient for clinical narratives. Clinical records frequently describe multiple medical events and treatment stages within the same document (Hazlehurst et al., 2005). As a result, surface-level semantic matching may retrieve sentences that share keywords or embeddings with the answer but refer to different clinical episodes

### 1. Clinical episode identification

locate the medical event referenced in the answer. (e.g., surgery, examination, or hospitalisation).

### 2. Stage of care determination

determine the stage of the treatment process the answer describes.

background → finding → decision → procedure → outcome.

### 3. Evidence retrieval

search the clinical note for sentences describing the same episode and stage.

### 4. Evidence verification

ensure that each clinical fact in the answer is supported by at least one cited sentence.

Figure 1: Four-step clinical reasoning principles declaration, utilised as a part of our zero-shot learning methodology.

or stages of care.

Recent studies have explored the use of large language models (LLMs) and prompting techniques to support clinical tasks. For example, prompt engineering guidelines have been proposed to help clinicians design effective prompts by defining clinical objectives, following medical reasoning principles, and evaluating prompt quality (Liu et al., 2025). Other work investigates instruction tun-



ing and prompt optimisation strategies to improve LLM performance in medical applications (Le et al., 2025; Dhamija and Sharma, 2025).

Motivated by these developments, we propose a prompt-based evidence-alignment method that leverages large language models and incorporates principles of clinical reasoning, using GPT 5.2 by OpenAI. Instead of relying on surface-level similarity, our approach guides the model to identify the clinical episode and stage of care referenced in each answer sentence. The clinical reasoning principles are shown in Figure 1. Experimental results on ArchEHR-QA show that incorporating structured clinical reasoning improves the evidence alignment performance.

## 2. Methodology

**Task Definition:**<sup>1</sup> Given a case  $(Q_c, E, A)$ , where  $Q_c$  is clinician-interpreted question,  $E = \{e_1, e_2, \dots, e_m\}$  is a *clinical note excerpt* that consist of sentences  $e_i, i \in \overline{1, m}$ , and  $A = \{a_1, a_2, \dots, a_n\}$  is an *answer* that yields of sentences  $a_i, i \in \overline{1, n}$ . For each  $a_i \in A$ , the task is to identify supporting sentences ( $E_s$ ), where  $E_s \subseteq E$ .

Our methodology represents prompt-based engineering, with the structure of the *clinical prompt* depicted in Figure 2, with the details on the prompt components covered in Figure 3. It features **clinical reasoning principles** which are in greater detail covered in Section 2.1.

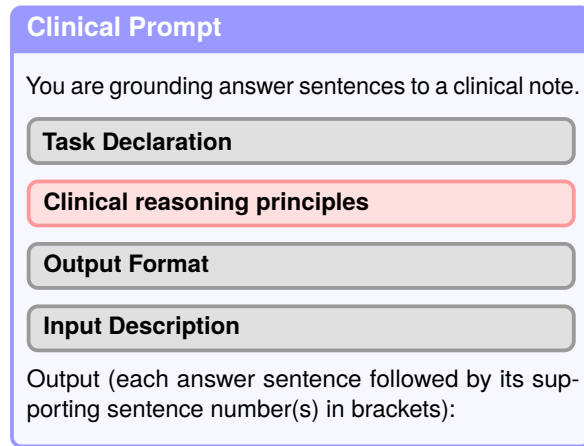


Figure 2: Clinical Prompt which features clinical reasoning principles (see Section 2.1)

For the given input example  $(Q_c, E, A)$ , we compose prompt to query the model to align the whole answer ( $A$ ) in zero-shot<sup>2</sup> learning mode.

<sup>1</sup>Our methodology omits patient question.

<sup>2</sup>Our approach could be viewed as “one-shot” since *Output Format* includes a single-sentence example (Figure 3); however the intention for including the example is to propose the expected output format for the task

### Task Declaration

#### Goal:

For each answer sentence, identify the clinical note sentence(s) that support it.

#### Input:

Clinician-interpreted question (Clinician-Interpreted Question)  
Clinical note excerpt with numbered sentences (Clinical Note Excerpt)  
Answer text to be grounded (Reference Answer)

#### Expected output:

For each answer sentence, a set of supporting evidence sentence numbers from the clinical note excerpt.

### Clinical reasoning principles

Clinical documentation follows a progression:  
background → finding → decision → procedure → outcome.

For each answer sentence:

#### List of Rules

If any clinical fact is unsupported, search again and add the necessary sentence. Do not output this verification step.

### Output Format

#### Output format:

Output each answer sentence followed by its supporting sentence number(s) in brackets.

**Example:** "The patient had elevated bilirubin [3]. ERCP was performed [5]."

### Input Description

---

#### Clinician-Interpreted Question:

$Q_c$

#### Clinical Note Excerpt (numbered sentences):

$E$

#### Reference Answer (ground each sentence to note sentences above):

$A$

---

Figure 3: Components of the clinical prompt and with content for “list of rules”, described in Section 2.1; for the *Output Format*, we include a fixed “example” for leveraging in-context learning; bold is used for highlighting the key words that are used in the prompt.

**Structuring the output.** To structure the raw output, we applied the following steps: (i) identifying

list of sentences/chunks (ii) mapping the most likely relevant sentence for the chunk.

We use “\n” char to demarcate content into sentences. We apply regular expression to retrieve content in “[ ]” by covering: separated entries, dash-separated entries (ranges). We use the Levenshtein distance (Levenshtein et al., 1966) to determine the best match between the output chunk and the sentence.

## 2.1. Clinical Reasoning Principles Prompt Design

Clinical records typically document the progression of a patient’s treatment, including clinical findings, treatment decisions, procedures performed, and outcomes. In evidence-based medicine, clinicians interpret treatment conclusions by associating clinical observations with specific medical events and stages of care (Zakowski et al., 2004).

Inspired by this process, we incorporate principles of clinical reasoning into prompt design to guide evidence alignment. Specifically, the model needs to reason according to the following steps:

**Step 1: Clinical episode identification.** The model first identifies the clinical episode referenced by the answer sentence, such as a specific surgery, examination, or hospitalisation. Since clinical terms could refer to multiple episodes, it is easy to misalign the answer with sentences describing other events if events are not distinguished. Therefore, we explicitly require the model in the prompt to only look for evidence within the scope of the corresponding clinical episode.

**Step 2: Stage-of-care identification.** After identifying the episode, the model needs to determine which stage of the treatment process the answer sentence describes. We categorise common treatment stages in clinical records into four types: clinical findings, treatment decisions or recommendations, procedures performed, and outcomes. The model needs to determine the stage to which the answer sentence belongs based on its semantics, then search for sentences in the clinical records that describe the same event and stage as evidence.

**Step 3: Evidence selection rules.** To ensure reliable grounding, we designed several rules to strengthen the principle and avoid misalignment.

- Sentences that clearly show that clinical events have happened should be given first priority. For example, statements that say surgery has been done or a diagnostic result has been seen.
- Avoid selecting sentences that only share keywords with the answer but talk about clinical events or background information that isn’t relevant.

### List of Rules

1. Identify which clinical episode it refers to (e.g., first ERCP, repeat ERCP, specific hospital day). Only search within that episode.
2. Determine what stage of care the answer sentence represents:
  - a clinical finding
  - a decision/recommendation
  - a performed procedure
  - a result/outcome
3. Select the sentence(s) in the clinical note that directly document that specific stage of care.
4. Prefer explicit documentation of completed events (e.g., “was performed”, “showed”, “was found”) rather than related background or earlier mentions.
5. Do NOT cite sentences:
  - from a different episode
  - that only share keywords
  - that describe a different stage of care
6. Cite the minimal number of sentences needed. Only cite a sentence if removing it would make the answer unsupported.
7. Before finalising the citation set for an answer sentence, verify that every distinct clinical fact expressed in the sentence is explicitly supported by at least one cited clinical note sentence.

Figure 4: List of Rules that a part of clinical reasoning principles

- Select the minimum set of sentences required to support the answer statement.

Figure 4 lists evidence selection rules.

**Step 4: Evidence Verification.** Finally, we added a validation step that requires the model to verify that each clinical fact expressed in the answer sentence is supported by at least one citation from a clinical record, and to check the answer before giving the final answer.

## 3. Dataset

We use the officially provided ArchEHR-QA dataset ( $D$ ) version 1.5 (Soni and Demner-Fushman, 2026a). Table 2 presents a verbose analysis of cases for excerpt and answer lengths, separately for development ( $D_{dev}$ ) and test ( $D_{test}$ ) splits. According to the related comparison, with nearly similar length of sentences in excerpts and answers, **answers are  $\approx 5.32$  times shorter** (in sentences) than excerpts.

| Method                              | $F1_{overall}$ | $P_{micro}$ | $R_{micro}$ | $F1_{micro}$ | $P_{macro}$ | $R_{macro}$ | $F1_{macro}$ |
|-------------------------------------|----------------|-------------|-------------|--------------|-------------|-------------|--------------|
| $D_{dev}$                           |                |             |             |              |             |             |              |
| LLaMA-3-70B-instruct ( $\ddagger$ ) | 67.95          | 68.28       | 72.26       | 70.21        | 75.51       | 76.29       | 65.68        |
| GPT-5.2 ( $\dagger$ )               | 74.77          | 70.69       | 89.13       | 78.85        | 73.01       | 90.50       | 70.68        |
| GPT-5.2 ( $\ddagger$ )              | 76.24          | 77.48       | 84.78       | 80.97        | 79.39       | 86.87       | 71.51        |
| $D_{test}$                          |                |             |             |              |             |             |              |
| GPT-5.2 ( $\ddagger$ ) (ours)       | 77.41          | 75.70       | 77.38       | 76.53        | 79.28       | 80.81       | 78.29        |
| team boris123 (baseline)            | —              | 76.50       | 58.20       | 66.10        | —           | —           | —            |

Table 1: Results on ArchEHR-QA separately for  $D_{dev}$  and  $D_{test}$  splits;  $\dagger$  - excluding reasoning principles from the prompt,  $\ddagger$  - including reasoning principles (Section 2.1).

| Parameters                     | $D_{dev}$ | $D_{test}$ |
|--------------------------------|-----------|------------|
| Cases (Total)                  | 20        | 147        |
| Excerpt (chars / sentence avg) | 95.98     | 93.13      |
| Excerpt (sentences avg)        | 21.4      | 28.5       |
| Excerpt (sentences min)        | 9         | 5          |
| Excerpt (sentences max)        | 54        | 73         |
| Answer (chars / sentence avg)  | 102.9     | 115.2      |
| Answer (sentences avg)         | 4.8       | 4.6        |
| Answer (sentences min)         | 3         | 2          |
| Answer (sentences max)         | 6         | 8          |
| Answer (cites avg)             | 6.9       | —          |
| Answer (cites / sentence avg)  | 1.5       | —          |

Table 2: Parameter comparison for ArchEHR-QA.

## 4. Experimental Setup

We test our approach on the proprietary GPT-5.2 model <sup>2025-12-11</sup><sup>3</sup> from OpenAI while including LLaMA-3-70B as a comparison. Our primary focus is on GPT-5.2, and we do not perform an in-depth analysis of LLaMA-3-70B. We set the temperature to 0.1 for both models. The prompts that follows the methodology outlined in Section 2 are publicly available in repository <sup>4</sup>.

To assess the impact of clinical reasoning principles, we experiment with prompt variations: excluding reasoning principles from the prompt ( $\dagger$ ), and including reasoning principles ( $\ddagger$ ).

**Evaluation metrics:** The ArchEHR-QA 2026 organizers chose  $F1_{micro}$  and  $F1_{macro}$  as evaluation metrics. We found no publicly available evaluation script. For the results on  $D_{dev}$ , we use manually implemented calculation of  $F1_{micro}$  and  $F1_{macro}$ . To obtain the related results, we first calculate the related  $F1$  values at the sentence level and then aggregate them to the case level. The results for  $D_{test}$  split are obtained from the official evaluation script.

<sup>3</sup><https://openai.com/index/introducing-gpt-5-2/>

<sup>4</sup>[https://github.com/nicolay-r/ArchEHR-QA-2026-Task-4-MedEvi-NS/blob/master/src/utils\\_prompt.py](https://github.com/nicolay-r/ArchEHR-QA-2026-Task-4-MedEvi-NS/blob/master/src/utils_prompt.py)

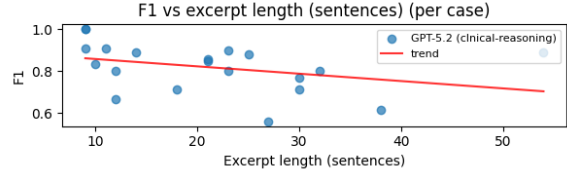


Figure 5: Detailed overview  $F1_{overall}$  results for GPT-5.2 ( $\ddagger$ ) on  $D_{dev}$  based on the length of the excerpt notes in sentences

## 5. Result Analysis and Discussion

Table 1 presents results on ArchEHR-QA. We include results of the open-sourced model LLaMA-3-70B-instruct<sup>5</sup> for baseline purposes. We found that LLaMA-3-70B underperformed compared to GPT-5.2. Furthermore, we found no improvements in adopting clinical reasoning principles for LLaMA-3-70B (Uzun Bektaş et al., 2025).

Towards results on  $D_{dev}$ , adopting clinical reasoning principles ( $\ddagger$ ) for GPT-5.2 results in +2.0% improvement by  $F1_{overall}$  on  $D_{dev}$ , leveraged by increased precision. Therefore, for our final submission on  $D_{test}$  we adopted GPT-5.2  $\ddagger$ .

Towards results on  $D_{test}$  split, we refer to the results of team “boris123” as the baseline. Using GPT-5.2  $\ddagger$  improves the baseline +15.8% ( $F1_{micro}$ ). Notably, with almost similar  $P_{micro}$  results, our system finds more citations.

According to findings from analysis of ArchEHR-QA content dataset in Section 5, *clinical excerpts* represent a significant portion once imputed in Input Description (Figure 3). To find whether the length of excerpts affects the performance, we provide case-level evaluation for our results on  $D_{dev}$  split in Figure 5. According to the related results, we found that the increment of excerpt from 10 sentences to 50 causes  $\approx 10\%$  degradation of  $F1_{overall}$  performance for GPT-5.2.

<sup>5</sup><https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct>

## 6. Conclusion

This paper proposes a clinical-reasoning principle, utilised in a form of prompt-based engineering methodology for ArchEHR-QA 2026 Subtask 4. For the given clinical interpretation of the patient question ( $Q_c$ ), note excerpt ( $E$ ), and answer text ( $A$ ), this subtask aims to cite supporting sentences from  $E$  to sentences from  $A$ . We propose a prompt-based engineering methodology that exploits clinical reasoning principles to leverage LLM capabilities for zero-shot learning on both a proprietary model (GPT-5.2) and an open-sourced model (LLaMA-3-70B). According to our experiment on the officially provided ArchEHR-QA dataset, we found that applying clinical-reasoning principles improves  $F1_{overall}$  for GPT-5.2 by +2.0%. We see those findings as a promising step for future works on the formation of advanced systems that utilise the Chain-of-Thought (CoT) paradigm.

## 7. Bibliographical References

- Akanksha Dhamija and Arun Sharma. 2025. Prompt-enhanced question answering for clinical texts using instruction-tuned language models. In *2025 1st IEEE Uttar Pradesh Section Women in Engineering International Conference on Electrical Electronics and Computer Engineering (UPWIECON)*, pages 308–312. IEEE.
- Shashi Kant Gupta, Aditya Basu, Mauro Nievas, Jerrin Thomas, Nathan Wolfrath, Adhitya Ramamurthi, Bradley Taylor, Anai N Kothari, Regina Schwind, Therica M Miller, et al. 2024. Prism: Patient records interpretation for semantic clinical trial matching using large language models. *arXiv preprint arXiv:2404.15549*.
- Brian Hazlehurst, H Robert Frost, Dean F Sittig, and Victor J Stevens. 2005. Mediclass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. *Journal of the American Medical Informatics Association*, 12(5):517–529.
- K Karthik, Sowmya Kamath, R Supreetha, and Ashish Katlam. 2025. Content-based medical retrieval systems with evidence-based diagnosis for enhanced clinical decision support. *Expert Systems with Applications*, 272:126678.
- Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Chanwoo Park, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, et al. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.
- Chenqian Le, Ziheng Gong, Chihang Wang, Haowei Ni, Panfeng Li, and Xupeng Chen. 2025. Instruction tuning and cot prompting for contextual medical qa with llms. In *2025 International Conference on Artificial Intelligence, Human-Computer Interaction and Natural Language Processing (ICAHN)*, pages 43–46. IEEE.
- Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.
- Jialin Liu, Fang Liu, Changyu Wang, and Siru Liu. 2025. Prompt engineering in clinical practice: tutorial for clinicians. *Journal of Medical Internet Research*, 27:e72644.
- Kathleen N Lohr, Kristen Eleazer, and Josephine Mauskopf. 1998. Health policy issues and applications for evidence-based medicine and clinical practice guidelines. *Health policy*, 46(1):1–19.
- 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.
- Sarvesh Soni and Dina Demner-Fushman. 2026a. [A dataset for addressing patient’s information needs related to clinical course of hospitalization. Scientific Data.](#)
- 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.
- Aslıhan Uzun Bektaş, Balahan Bora, and Erbil Ünsal. 2025. Comparative evaluation of chatgpt and llama for reliability, quality, and accuracy in familial mediterranean fever. *European Journal of Pediatrics*, 184(8):491.
- Laura Zakowski, Christine Seibert, and Wisconsin Selma VanEyck. 2004. Evidence-based medicine: answering questions of diagnosis. *Clinical medicine & research*, 2(1):63–69.
- Xinping Zhao, Dongfang Li, Yan Zhong, Boren Hu, Yibin Chen, Baotian Hu, and Min Zhang. 2024. Seer: Self-aligned evidence extraction for retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3027–3041.



# UIC-AIHealth4All at ArchEHR-QA 2026: Answer-First Evidence Grounding for Clinical Question Answering

Mohammad Arvan, Hossein Haeri, Natalie Parde, Rebecca Feinstein

University of Illinois Chicago  
Chicago, IL, USA  
{marvan3, haeri, parde, rfeinst}@uic.edu

## Abstract

We describe the UIC-AIHealth4All system for ArchEHR-QA 2026, a shared task on grounded question answering from electronic health records. We participated in Subtasks 2 (evidence identification), 3 (answer generation), and 4 (answer-evidence alignment). For Subtasks 2 and 3, we propose an *answer-first* pipeline in which the model generates candidate answers citing specific note sentences before classifying the full evidence set, exploiting the asymmetry between judging relevance in the abstract versus relative to a generated answer. For Subtask 4, we apply self-consistency voting over five independent model calls, retaining links by vote threshold. Our pipeline ranked third on evidence identification (Strict Micro F1 62.90), ninth on answer generation (Overall 31.90), and fifth on answer-evidence alignment (F1 79.81). A post-hoc linguistic analysis of 45 stylistic features reveals that model outputs remain 3.2 Flesch-Kincaid grade levels harder to read than clinician-authored references despite matching their word and sentence counts, suggesting readability warrants explicit optimization in clinical NLP systems. Code and prompts are available at <https://github.com/mo-arvan/archehr-qa-2026-uic-aihealth4all>.

**Keywords:** clinical NLP, question answering, electronic health records, evidence identification, grounded generation

## 1. Introduction

Grounding an answer means anchoring every claim to a specific passage in the source document. In safety-critical domains such as clinical question answering, grounding is a reliability requirement: an answer that cannot be verified against the source text cannot be trusted for medical decisions. This requirement creates a joint constraint on evidence selection and answer generation. The minimal evidence set depends on what the answer will say, and a faithful answer cannot exceed what the evidence supports.

Grounded QA from electronic health records (EHRs), where both operations must be solved jointly, has received less attention than individually scoped clinical QA tasks (Soni and Demner-Fushman, 2025). Prior work has addressed factoid retrieval from structured data (Lee et al., 2022), reading comprehension over published literature (Jin et al., 2019), and information extraction from clinical text (Uzuner et al., 2011), each formulated as an independent subtask. Whether generating a candidate answer before classifying evidence improves the accuracy of evidence identification has not been established.

ArchEHR-QA 2026 (Soni and Demner-Fushman, 2026b,a) directly addresses grounded EHR question answering across four subtasks: question interpretation (Subtask 1), evidence identification (Subtask 2), answer generation (Subtask 3), and answer-evidence alignment (Subtask 4). The dataset is derived from clinical notes in the MIMIC-III (Johnson et al., 2016) and

MIMIC-IV (Johnson et al., 2023) databases, with patient-authored questions, clinician-interpreted reformulations, sentence-level relevance annotations, and clinician-authored reference answers with sentence-level evidence alignments. The shared task provides a setting in which all three components of the grounded QA pipeline (evidence, answer, and alignment) can be evaluated against expert clinical judgments.

We participated in three of the four subtasks (Subtasks 2, 3, and 4) as the UIC-AIHealth4All team. Our central contribution is an answer-first pipeline for evidence identification and answer generation (Subtasks 2 and 3), extending our prior system (Arvan et al., 2025), in which the model first generates one or more candidate answers citing specific note sentences, then classifies the full evidence set in light of those candidates, and finally produces a submission answer from the classified evidence. For Subtask 4, we applied self-consistency voting over multiple independent model calls, aggregating link predictions by vote count and selecting a threshold that balanced precision and recall on the development set.

We additionally characterized stylistic differences between model outputs and clinician-authored reference answers. We computed 45 linguistic features across six categories (lexical, syntactic, readability, stylistic, clinical, and sentiment) and compared them with paired Wilcoxon signed-rank tests. This linguistic analysis guided prompt refinement and revealed a residual register gap between model outputs and clinician answers that persists after length and sentence count are matched.



Our main results are as follows: (1) the answer-first pipeline ranked third on evidence identification (Strict Micro F1 62.90); (2) our answer generation submission ranked ninth overall (test score 31.90), with iterative constraint enforcement achieving 100% output compliance on both development and test sets within five correction attempts; (3) self-consistency voting ranked fifth on answer-evidence alignment (F1 79.81); and (4) despite matching clinician word count and sentence count, model outputs remained 3.2 Flesch-Kincaid grade levels harder to read than the clinician reference. These results suggest that answer-first ordering offers a practical advantage for evidence identification and that readability warrants explicit optimization in clinical NLP systems serving lay users.

## 2. Method

### 2.1. Study Design and Setting

We participated in Subtasks 2, 3, and 4 of ArchEHR-QA 2026 (Soni and Demner-Fushman, 2026b), a shared task on grounded question answering from clinical notes. Subtask 2 (evidence identification) asks a system to select the minimal set of note sentences that support answering the patient question. Subtask 3 (answer generation) asks a system to generate a natural-language answer of at most 75 words, grounded in those sentences. Subtask 4 (answer-evidence alignment) asks a system to map each sentence of a pre-written reference answer to the specific note sentences that support it.

### 2.2. Data

All experiments used the ArchEHR-QA 2026 development and test sets, derived from the MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023) clinical databases. Each case contains a patient-authored question, a clinician-interpreted reformulation (at most 15 words), a clinical note excerpt with sentence-level relevance annotations, and a clinician-authored reference answer with sentence-level evidence alignments. The development set contains 20 cases. The test set for Subtasks 2 and 3 contains 47 cases; the test set for Subtask 4 contains 147 cases.

### 2.3. Systems

We designed two systems building on our prior work (Arvan et al., 2025): an answer-first pipeline for Subtasks 2 and 3, and a self-consistency voting pipeline for Subtask 4. Code and prompts are available at <https://github.com/mo-arvan/archehr-qa-2026-uic-aihealth4all>.

All experiments used GPT-5.1 and GPT-5.2 through the Azure OpenAI API with structured out-

puts (`responses.parse`) and Pydantic schema validation. Both models support a reasoning effort parameter; we swept over *medium* and *high* settings during development and used *medium* for all test submissions. The Subtask 2 ensemble spanned both model versions and multiple  $k$  settings; Subtask 4 used GPT-5.2. No local GPU compute was required.

**Answer-first pipeline (Subtasks 2 and 3).** We designed a modular pipeline that jointly produced evidence predictions for Subtask 2 and an answer for Subtask 3 in three sequential stages.

**Stage 1: candidate answer generation.** We prompted the model to generate  $k$  independent candidate answers, each citing specific note sentences by ID. The full note, the patient question, and the clinician-interpreted question were provided as input. The generation prompt specified past tense for events during the hospitalization and present tense for current state and prognosis, required third-person subject constructions (“The patient”), and targeted a word count of 70 to 75 words.

**Stage 2: answer-grounded evidence classification.** We presented the  $k$  candidates alongside the clinical note and instructed the model to classify each note sentence as *Essential*, *Supplementary*, or *Not Relevant*, using a minimal-set criterion: a sentence was *Essential* only if removing it would leave the generated answers unsupported. We refer to this method as the *answer-first* pipeline. It exploits an asymmetry in difficulty: judging which sentences the model relied upon after generating an answer is easier than judging relevance in the abstract.

**Stage 3: final answer generation.** We instructed the model to generate the submission answer using only *Essential* (and optionally *Supplementary*) sentences, subject to the 75-word constraint. Each response was validated against the word limit, citation consistency, and output format. Constraint violations triggered a correction call that named the specific error and included the previous response as context, with up to five attempts per case. We logged all constraint violations by error type across all experiments to characterize the dominant failure modes of the pipeline.

For our best Subtask 2 submission, we combined seven configurations by majority vote: a note sentence was predicted *Essential* if at least four of the seven runs classified it as such. Ensemble members spanned both model versions and multiple  $k$  settings to maximize prediction diversity.

**Self-consistency voting pipeline (Subtask 4).** We designed a self-consistency voting pipeline for Subtask 4 in which  $N = 5$  independent model calls per case were aggregated by vote count to produce the final set of answer-sentence to note-sentence link predictions.

**Prompt design.** The alignment prompt instructed the model to apply a precision-first principle: cite only the note sentence that directly states the specific fact in the answer sentence. The prompt documented four common alignment errors with corrective examples: (1) over-citing, where contextual sentences are cited instead of the direct-claim sentence; (2) under-citing, where distinct claims sharing a source sentence are omitted; (3) citing synthesized conclusions, where no single note sentence directly states the inference; and (4) citing structural text such as section headers or symptom checklists. A worked example from the task documentation illustrated correct and incorrect alignments for each error type.

**Self-consistency voting.** Each of the  $N = 5$  samples produced a set of (answer sentence, note sentence) link pairs. A link was included in the final prediction if it appeared in at least  $t$  of the 5 samples. We selected  $t = 3$  based on a precision-recall trade-off analysis on the development set.

## 2.4. Measures and Outcomes

We evaluated system outputs using the official task metrics and supplementary compliance measures.

For Subtask 2, gold labels classify each sentence as *essential*, *supplementary*, or *not-relevant*. We report precision, recall, and F1 against *essential* sentences (strict) and against the union of *essential* and *supplementary* (lenient). We additionally measured first-attempt constraint compliance as the percentage of outputs satisfying all format and length requirements on the first generation attempt, and logged all violations by error type.

For Subtask 3, we report lexical overlap (BLEU, ROUGE), semantic similarity (BERTScore, AlignScore), medical concept coverage (MEDCON), and simplification quality (SARI).

For Subtask 4, we report precision, recall, and F1 over predicted alignment links.

## 2.5. Statistical and Analytic Methods

To characterize how closely model outputs matched the stylistic register of clinician-authored reference answers, we computed 45 linguistic features for each of the 20 development cases and compared them against the clinician references. Features spanned six categories: lexical (word count, type-token ratio, Measure of Textual Lexical Diversity (MTLD), hapax ratio, average word length), syntactic (sentence count, average sentence length, parse depth, part-of-speech (POS) ratios), readability (Flesch-Kincaid Grade, Coleman-Liau Index, Flesch Reading Ease), stylistic (function-word percentage, passive-voice ratio, pronoun usage, frequency of the phrase *the patient*), sentiment

and tone (polarity, subjectivity, hedging and warning counts), and clinical communication (certainty markers, temporal and causal connectives, medical abbreviation count).

We applied paired Wilcoxon signed-rank tests to identify features that systematically diverged between model and clinician outputs. This non-parametric test was chosen because the 20-case sample cannot be assumed to yield normally distributed feature differences, and because each case contributes a matched model-clinician pair. We additionally computed Cohen’s  $d$  for each feature, a standardized effect size that expresses the mean paired difference in units of the within-case standard deviation, to distinguish statistically significant differences from practically large ones. To summarize overall stylistic proximity, we defined a standardized linguistic distance as the mean of  $|\text{model} - \text{clinician}| / \text{SD}_{\text{clinician}}$  across all features, where  $\text{SD}_{\text{clinician}}$  is the between-case standard deviation of the clinician values.

We applied this analysis to two configurations: the submitted two-step pipeline and the exploratory three-step rewrite variant. To visualize category-level profiles, we produced a radar chart with six composite scores, one per feature category, normalized so that the clinician profile traces a unit circle (Figure 2). Each spoke value is the mean of per-feature model-to-clinician ratios within that category. Values above 1.0 indicate that the model exceeds the clinician on that dimension; values below 1.0 indicate that it falls short.

## 2.6. Sensitivity Analyses

**Task decomposition ablation.** The linguistic analysis revealed a stylistic gap between model outputs and clinician-authored answers, suggesting that generating factually grounded, stylistically appropriate, and length-constrained text within a single prompt creates competing objectives. To test whether separating these objectives narrows the gap, we compared the submitted pipeline against a three-stage *rewrite* variant that (1) generates an unconstrained answer, (2) rewrites it for clinical register, and (3) enforces the 75-word limit as a separate call. The rewrite variant was implemented after the submission deadline. We evaluated word count compliance and standardized linguistic distance to the clinician reference as outcome measures.

# 3. Results

## 3.1. Subtask 2: Evidence Identification

The answer-first ensemble achieved Strict Micro F1 62.90 on the test set. On the development set, the best single-run configuration achieved F1 66.22.

For the test submission, we combined seven configurations by majority vote, predicting a sentence as Essential if at least four of seven runs classified it as such. Table 1 shows development and test results.

The ensemble outperformed the best individual test submission by 3.2 F1 points. Development recall (81.82%) exceeded test recall (66.96%) by 14.9 percentage points, suggesting distribution shift on the larger 47-case test set.

### 3.2. Subtask 3: Answer Generation

The answer-first pipeline achieved test Overall 31.90 (development Overall 32.21). Table 2 shows full results for the submitted configuration.

SARI was the highest-scoring metric on both sets (58.77 development, 57.24 test), reflecting simplification quality relative to the source note. BLEU was the lowest (4.01 development, 6.12 test), consistent with the observation that model outputs differed in surface form from the reference answers even when semantic similarity (BERTScore 41.89, AlignScore 25.00) and concept coverage (MEDCON 41.30) were moderate on the development set.

### 3.3. Subtask 4: Answer-Evidence Alignment

Self-consistency voting achieved test F1 79.81 on Subtask 4 (development F1 86.0). We drew  $N = 5$  independent samples per case and retained links appearing in at least  $t = 3$  of 5 samples. Table 3 shows development and test results.

The 6.2-point development-to-test F1 drop (86.0 to 79.8) reflects the larger and more diverse test set (147 cases, including 127 cases from the 2025 dataset). Across all five teams on the official leaderboard, precision ranged from 82 to 88 and recall from 71 to 78, indicating that recall was the primary differentiator in this task.

We selected  $N = 5$  and  $t = 3$  based on practical and empirical considerations. Five samples balanced prediction diversity against API cost, representing a  $5\times$  increase over a single call. Lowering  $t$  from 4 to 3 recovered 4.2 percentage points of recall at a cost of 2.4 points of precision, yielding a net F1 gain of 1.34 points on the development set.

### 3.4. Error Analysis

Iterative correction resolved all constraint violations across 40 experiments (1,171 total cases: 701 development, 470 test). The sections below characterize correction effectiveness and the distribution of failure modes.

#### 3.4.1. Correction Effectiveness

The correction loop achieved 100% compliance on both the development and test sets (see Appendix Table 5 for per-attempt rates). First-try compliance was 21.6 percentage points lower on the test set (64.3% versus 85.9%), reflecting longer clinical notes and unseen question types. The correction loop closed this gap fully within five attempts in both conditions.

#### 3.4.2. Error Taxonomy

Word count violations accounted for 75.2% of development errors and 94.9% of test errors. Citation errors, schema errors, and pipeline failures were absent from the test set entirely, indicating that development-set tuning eliminated structural and citation errors but did not generalize to length constraint adherence on longer test notes. Figure 1 shows the full breakdown.

#### 3.4.3. Content Safety

Two distinct content safety failure modes affected a small fraction of cases. Infrastructure-level content filtering (non-deterministic): the API silently returned no response, causing a downstream error. Eleven unique cases were affected (4 development, 7 test); retries typically succeeded. Model refusal (persistent): the model returned safety-decline text that could not be parsed as a structured response. Eight unique cases were affected (7 development, 1 test), all involving sensitive clinical content. Because these events were input-dependent, they could not be resolved by correction prompts; multi-candidate pipelines ( $k > 1$ ) provided partial resilience by allowing other candidates to succeed even when one call was refused.

### 3.5. Linguistic Analysis

#### 3.5.1. Model versus Clinician Divergence

Model outputs diverged from clinician-authored reference answers on 8 of 45 linguistic features (Wilcoxon signed-rank, Benjamini-Hochberg adjusted  $p < 0.05$ ). Table 6 (Appendix) lists all features for the submitted configuration, sorted by effect size (Cohen's  $d$ ).

Three divergence patterns dominated. First, model outputs were harder to read across all three readability metrics. The Flesch Reading Ease score was 21 points lower (46.7 versus 25.6,  $d = 1.11$ ), the Flesch-Kincaid grade was 3.2 points higher (10.7 versus 13.9,  $d = -0.93$ ), and the Coleman-Liau grade was 4.1 points higher (12.8 versus 16.9,  $d = -1.12$ ). Second, models used longer individual words (5.39 versus 6.12 characters,  $d = -1.21$ ) and fewer function words (40.9%

|            | St. P | St. R | St. F1       | Le. P | Le. R | Le. F1 |
|------------|-------|-------|--------------|-------|-------|--------|
| Dev (best) | 55.62 | 81.82 | 66.22        | 62.66 | 81.82 | 70.97  |
| Test (4/7) | 59.30 | 66.96 | <b>62.90</b> | 74.27 | 66.96 | 70.43  |

Table 1: Subtask 2 results for the best configuration (development) and the majority-vote ensemble (test). St. = Strict, Le. = Lenient. All metrics are micro-averaged.

|      | AS    | BS    | BLEU | MC    | R-1   | R-2   | R-L   | SARI  | Ov.          |
|------|-------|-------|------|-------|-------|-------|-------|-------|--------------|
| Dev  | 25.00 | 41.89 | 4.01 | 41.30 | 39.17 | 11.24 | 22.29 | 58.77 | 32.21        |
| Test | 24.53 | 41.80 | 6.12 | 37.52 | 38.84 | 11.95 | 24.18 | 57.24 | <b>31.90</b> |

Table 2: Subtask 3 results (Overall = mean of all reported metrics). AS = AlignScore, BS = BERTScore, MC = MEDCON, R = ROUGE, Ov. = Overall.

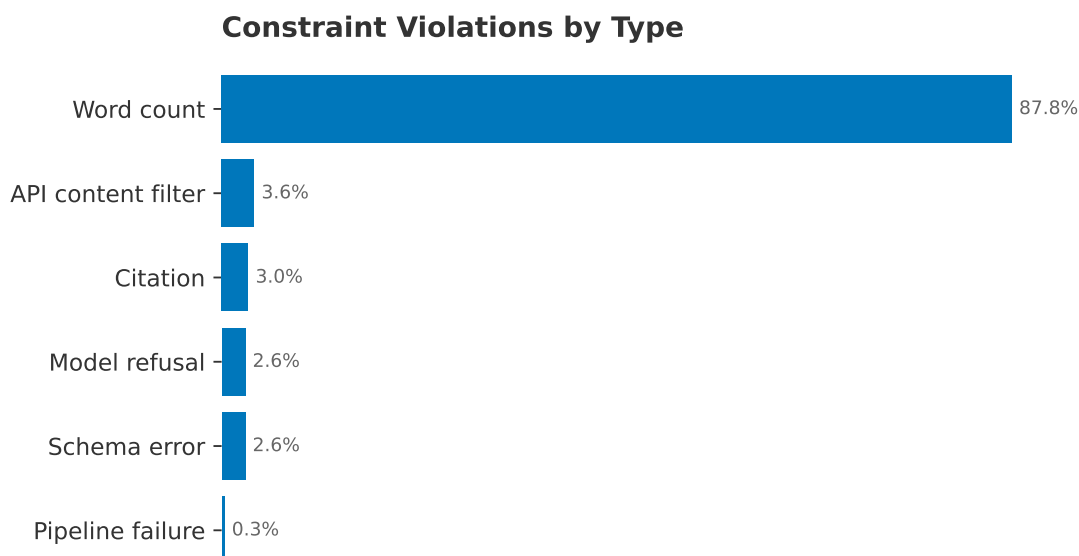


Figure 1: Distribution of first-attempt error types on the development ( $n = 129$  errors) and test ( $n = 177$  errors) sets. Word count violations dominated both conditions, rising from 75.2% of development errors to 94.9% of test errors.

|             | P           | R           | F1          |
|-------------|-------------|-------------|-------------|
| Development | 85.1        | 87.0        | 86.0        |
| Test        | <b>83.6</b> | <b>76.3</b> | <b>79.8</b> |

Table 3: Subtask 4 results (P = precision, R = recall, F1 = micro F1).

versus 34.2%,  $d = 1.26$ ), producing denser prose with higher noun and adjective rates. Third, models produced more lexically diverse outputs than clinicians, with higher type-token ratios (0.756 versus 0.791,  $d = -0.60$ ) and higher hapax ratios (0.599 versus 0.670,  $d = -0.77$ ).

Notably, word count (73.5 versus 72.7,  $p = 0.085$ ) and sentence count (4.75 versus 4.45,  $p = 0.303$ ) did not differ significantly. The submitted configuration brought word count and sentence count within the range of clinician values, resolving the length and sentence structure divergences seen in earlier configurations.

### 3.5.2. Two-Step versus Three-Step Linguistic Distance

The three-step rewrite pipeline reduced linguistic distance from 0.519 to 0.381 and improved Subtask 3 Overall by approximately 2 points on the development set. Table 4 shows the full comparison; Figure 2 shows category-level profiles as a radar chart normalized to the clinician reference.

Improvements were distributed across all six feature categories. The most pronounced gains were in readability (Flesch-Kincaid grade 12.52 versus 13.87, closer to the clinician's 10.72) and syntactic structure (sentence count 4.65 versus 4.45; average sentence length 19.17 versus 19.88 words). Passive voice ratio remained below the clinician reference in both configurations, and “the patient” frequency was below clinician levels in both (0.95 for Three-Step, 1.10 for Two-Step, versus clinician 1.55). The Three-Step pipeline achieved closer stylistic alignment and also improved Subtask 3 au-

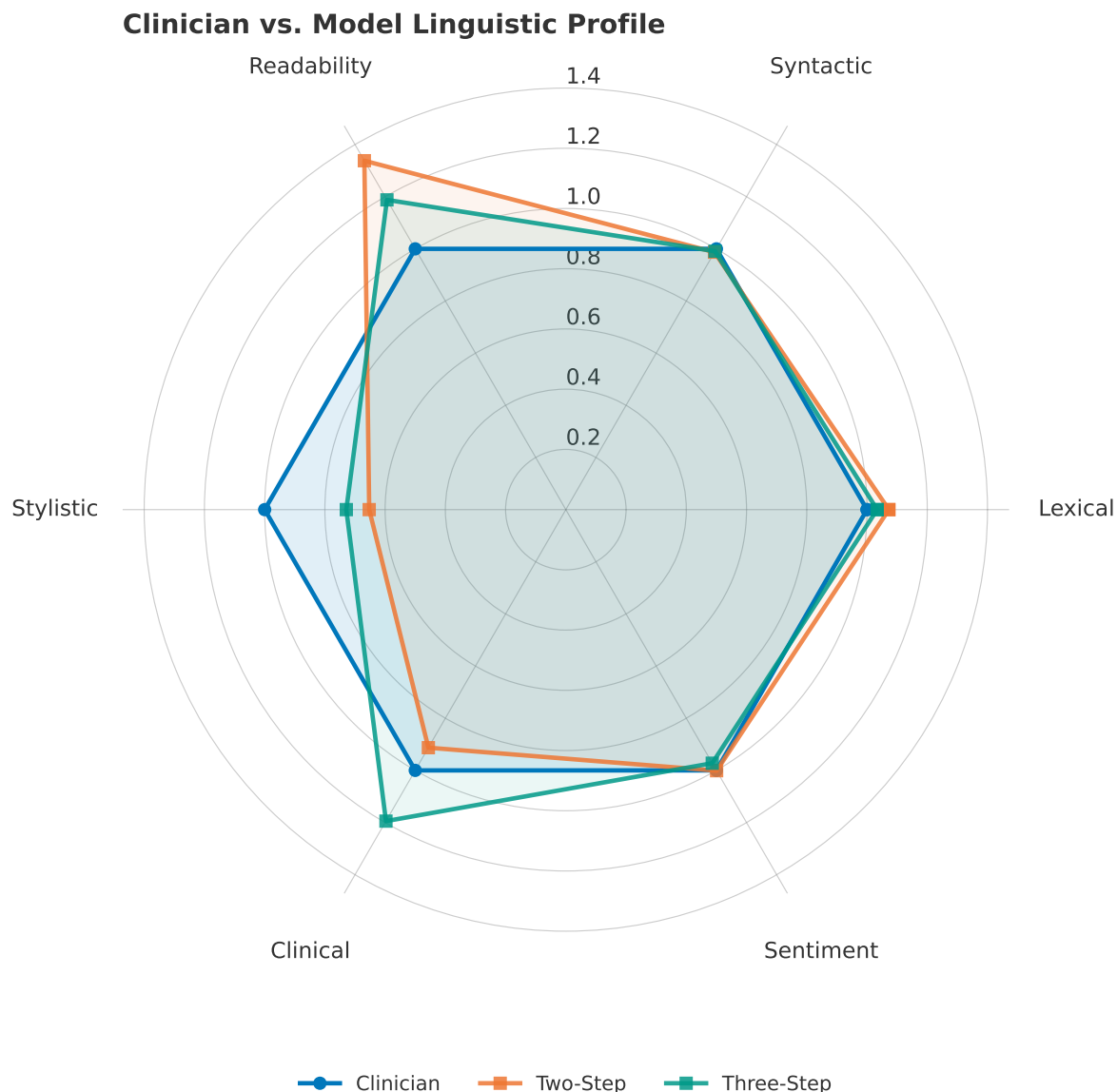


Figure 2: Category-level linguistic profiles for the Two-Step and Three-Step configurations. The clinician reference traces the unit circle. Spoke values are the mean of per-feature model-to-clinician ratios within each category, with all features weighted equally; values above 1.0 indicate the model exceeds the clinician on that dimension, and values below 1.0 indicate the model falls short.

tomatic metrics: Dev Overall 34.09 versus 32.21 for the Two-Step configuration, a gain of approximately 2 points.

## 4. Discussion

### 4.1. Main Findings

LLMs equipped with answer-grounded evidence classification, iterative constraint enforcement, and self-consistency voting can achieve competitive performance on grounded EHR question answering. These results demonstrate that the answer-first paradigm improved both evidence selection

accuracy and output compliance compared to direct classification baselines. Iterative correction resolved all remaining constraint violations in our experiments. At the same time, our linguistic analysis reveals that even configurations that perform competitively on automatic metrics produce text that is measurably harder to read and stylistically less similar to clinician-authored answers than word count and sentence count alone would suggest.

### 4.2. Evidence Identification (Subtask 2)

The ensemble submission ranked third on evidence identification (Strict Micro F1 62.90), gaining ap-



| Feature              | Clinician | 3-Step       | 2-Step |
|----------------------|-----------|--------------|--------|
| Linguistic distance  | –         | <b>0.381</b> | 0.519  |
| Dev Overall (ST3)    | –         | <b>34.09</b> | 32.21  |
| Word count           | 73.5      | 73.4         | 72.7   |
| Sentence count       | 4.75      | 4.65         | 4.45   |
| Avg sentence length  | 17.98     | 19.17        | 19.88  |
| Flesch-Kincaid grade | 10.72     | 12.52        | 13.87  |
| Passive voice ratio  | 0.39      | 0.29         | 0.24   |
| “the patient” count  | 1.55      | 0.95         | 1.10   |
| Total hedging        | 0.50      | 0.45         | 0.40   |
| Temporal connectives | 0.35      | 0.60         | 0.80   |

Table 4: Linguistic distance and selected features for the Two-Step (submitted) and Three-Step (post-deadline) configurations. ST3 = Subtask 3.

proximately 3 F1 points over the best single-run submission (59.69). These gains support the central hypothesis that evidence classification is easier after generating a candidate answer than before: the model’s own prior generation provides an explicit context for judging which sentences were relied upon, reducing the difficulty of the minimal-set determination.

The test F1 of 62.90 places our system within a competitive range. Our best submission achieved higher recall than precision ( $P = 59.30$ ,  $R = 66.96$ ), consistent with the observation that under-retrieval is more costly than over-retrieval in this task.

The configuration sweep suggested that pipeline method had a larger effect on F1 than model version alone. However, reasoning effort and candidate count ( $k$ ) varied across runs, preventing clean isolation of individual factor contributions. Ensemble diversity, spanning both model versions and multiple  $k$  settings, was the primary driver of the test-set gain over single-run baselines.

#### 4.3. Answer Generation (Subtask 3)

The Subtask 3 submission achieved an overall score of 31.90. Interpreting this result requires caution because the official overall score aggregates nine heterogeneous metrics (lexical overlap, semantic similarity, medical concept coverage, and simplification quality) that may partially conflict. SARI was the highest-scoring individual metric (57.24), reflecting simplification quality relative to the clinical note source.

The linguistic analysis in Section 3 reveals an important gap between metric performance and stylistic register. Although the submitted configuration eliminated statistically significant word count and sentence count divergences from the clinician reference, 8 of 45 linguistic features remained significantly different after Benjamini-Hochberg correction. Model outputs were consistently harder to read (Flesch-Kincaid grade 3.2 points higher), used longer individual words (6.12 versus 5.39 charac-

ters), and contained fewer function words (34.2% versus 40.9%). These properties reflect a tendency toward technical, noun-heavy prose that may be less accessible to lay readers despite matching the reference on length. Whether this stylistic divergence matters for clinical utility cannot be determined from automatic metrics alone and requires human or clinical evaluation.

#### 4.4. Answer-Evidence Alignment (Subtask 4)

Self-consistency voting with  $N = 5$  samples per case and a fixed threshold of  $t = 3$  achieved test F1 79.81 on Subtask 4. The vote-count decomposition on the development set showed a clear monotonic relationship between vote count and precision (approximately 80 to 90% at  $t = 5$ , approximately 55 to 70% at  $t = 3$ , approximately 15 to 20% at  $t = 1$ ), motivating the threshold choice. Lowering  $t$  from 4 to 3 recovered 4.2 percentage points of recall (72.1% to 76.3%) while reducing precision by 2.4 points (86.0% to 83.6%), yielding a net F1 gain of 1.34 points.

The full leaderboard pattern for Subtask 4 indicates that all five participating teams were precision-heavy ( $P = 82$  to  $88$ ,  $R = 71$  to  $78$ ), and the spread in recall largely determined the ranking. This pattern suggests that recall is the primary differentiator in the answer-evidence alignment task. The development-to-test drop (F1 86.0 to F1 79.8) reflects the larger and more diverse test set: 127 of the 147 test cases came from the 2025 challenge dataset.

## 5. Conclusion

We described the UIC-AIHealth4All system for ArchEHR-QA 2026, which participated in three subtasks of grounded question answering from electronic health records. Our primary contributions are an answer-first evidence classification pipeline for Subtasks 2 and 3, an iterative constraint enforcement loop that achieved full output compliance in our experiments, and self-consistency voting for Subtask 4 alignment.

The answer-first pipeline enabled an ensemble that ranked third on evidence identification (Strict Micro F1 62.90). The iterative correction loop achieved 100% constraint compliance on both the development and test sets within five attempts, demonstrating robustness to the distribution shift introduced by longer and more diverse test notes. Self-consistency voting with a fixed threshold of  $t = 3$  ranked fifth on answer-evidence alignment (F1 79.81), with vote count providing a reliable proxy for link precision.

The linguistic analysis reveals a persistent stylistic gap between model outputs and clinician-authored answers that is not captured by the official evaluation metrics. Despite prompt engineering that successfully matched clinician word count and sentence count, model outputs remained substantially harder to read, used longer and less functional vocabulary, and showed fewer pronouns and passive constructions than the clinician reference. The post-deadline three-step rewrite pipeline narrowed this gap (linguistic distance 0.381 versus 0.519 for the submitted two-step pipeline) at a 9.5-fold latency cost. This result suggests that readability and stylistic register are separable from factual accuracy and that decomposed prompting can address them, but at a computational cost that may not be acceptable in production systems.

Future work should include human and clinical evaluation to determine whether the residual stylistic divergences affect patient understanding or clinical utility. Applying the answer-first paradigm to other grounded retrieval tasks outside the clinical domain would test the generality of the core hypothesis. Developing efficient alternatives to multi-stage decomposition that reduce the latency cost of register correction remains an open problem. Finally, extending the self-consistency voting approach to evidence classification and answer generation—not only alignment—may further improve recall without sacrificing the precision gains achievable through threshold tuning.

## 6. Limitations

We conducted no human or clinical evaluation. Automatic metrics may not capture dimensions of answer quality most relevant to patients or clinicians, such as appropriateness, emotional tone, or clinical accuracy.

The multi-stage pipeline requires at least three sequential API calls per case. The three-step rewrite variant increases latency by a factor of 9.5, which may preclude deployment in time-sensitive clinical settings.

All experiments used a single dataset (ArchEHR-QA 2026) derived from the MIMIC-III and MIMIC-IV databases. Performance may not generalize to other EHR systems, note types, clinical specialties, or patient populations.

The system depends on proprietary models (GPT-5.1 and GPT-5.2 through Azure OpenAI). Results may not transfer to open-weight models, and API behavior may change across versions without notice.

The linguistic analysis was conducted on 20 development cases, limiting statistical power for features with small effect sizes. Results for the strongest divergences (readability metrics,  $d > 0.9$ )

are robust, but findings for features near the significance threshold should be interpreted with caution.

## 7. Ethical Considerations

All clinical notes were accessed through credentialed PhysioNet accounts under a signed data use agreement.<sup>1</sup> We developed the system using Azure OpenAI Services under a university account with a Business Associate Agreement (BAA). We applied for and received modified abuse monitoring, which disables human review of prompts and completions by the API provider. No protected health information was used during prompt development or tuning.

## 8. Acknowledgements

This work was supported by the AI.Health4All Center at the University of Illinois Chicago. Natalie Parde was also supported by the National Institutes of Health under Grants No. R41NR020667, 1R61DA057629-01A1, and 1R01AG091762-01 and the National Science Foundation under Grant No. 2125411 during this time. 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.

## 9. Bibliographical References

- Mohammad Arvan, Anuj Gautam, Mohan Zalake, and Karl M. Kochendorfer. 2025. [UIC at ArchEHR-QA 2025: Tri-step pipeline for reliable grounded medical question answering](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 110–117, Vienna, Austria. Association for Computational Linguistics.
- 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.
- Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. [MIMIC-IV. PhysioNet](#). Version 2.2.

<sup>1</sup><https://physionet.org/news/post/gpt-responsible-use>

Alistair Johnson, Tom Pollard, and Roger Mark. 2016. [MIMIC-III Clinical Database](#). *PhysioNet*. Version 1.4.

Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. 2022. [EHRSQL: A practical text-to-sql benchmark for electronic health records](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Sarvesh Soni and Dina Demner-Fushman. 2025. Overview of the archehr-qa 2025 shared task on grounded question answering from electronic health records. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Sarvesh Soni and Dina Demner-Fushman. 2026a. [A dataset for addressing patient’s information needs related to clinical course of hospitalization](#). *Scientific Data*.

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.

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.

## A. Correction Effectiveness

|                 | Dev ( $n = 701$ ) | Test ( $n = 470$ ) |
|-----------------|-------------------|--------------------|
| After attempt 1 | 85.9%             | 64.3%              |
| After attempt 2 | 90.9%             | 72.8%              |
| After attempt 3 | 96.1%             | 86.2%              |
| After attempt 4 | 98.1%             | 91.1%              |
| After attempt 5 | 99.4%             | 98.1%              |
| Final           | 100.0%            | 100.0%             |

Table 5: Cumulative constraint compliance by correction attempt across 35 development and 10 test experiments.

## B. Linguistic Feature Analysis

| Feature                     | Clinician | Model | d     | $p_{adj}$ |     |
|-----------------------------|-----------|-------|-------|-----------|-----|
| Function word %             | 40.9%     | 34.2% | 1.26  | <0.001    | *** |
| Average word length         | 5.39      | 6.12  | -1.21 | <0.001    | *** |
| Coleman-Liau grade          | 12.8      | 16.9  | -1.12 | <0.001    | *** |
| Flesch Reading Ease         | 46.7      | 25.6  | 1.11  | <0.001    | *** |
| Flesch-Kincaid grade        | 10.7      | 13.9  | -0.93 | 0.005     | **  |
| Hapax ratio                 | 0.599     | 0.670 | -0.77 | 0.022     | *   |
| Adjective rate              | 8.2%      | 10.7% | -0.75 | 0.018     | *   |
| Hapax count                 | 44.0      | 48.7  | -0.72 | 0.031     | *   |
| Pronoun rate                | 6.3%      | 3.4%  | 0.69  | 0.080     | ns  |
| Type-token ratio            | 0.756     | 0.791 | -0.60 | 0.085     | ns  |
| Temporal connectives        | 0.35      | 0.80  | -0.59 | 0.085     | ns  |
| Noun rate                   | 31.0%     | 34.6% | -0.56 | 0.080     | ns  |
| Passive voice ratio         | 0.39      | 0.24  | 0.55  | 0.085     | ns  |
| MTLD                        | 87.0      | 100.8 | -0.54 | 0.085     | ns  |
| Third-person pronouns       | 3.45      | 1.85  | 0.53  | 0.085     | ns  |
| Causal connectives          | 0.90      | 0.35  | 0.52  | 0.100     | ns  |
| Medical abbreviations       | 0.95      | 1.65  | -0.51 | 0.087     | ns  |
| Unique word count           | 55.5      | 57.5  | -0.49 | 0.103     | ns  |
| First-person pronouns       | 0.15      | 0.00  | 0.41  | n/a       | n/a |
| Avg sentence length (words) | 17.98     | 19.88 | -0.38 | 0.142     | ns  |
| Sentence length SD          | 6.52      | 5.14  | 0.37  | 0.255     | ns  |
| Word count                  | 73.5      | 72.7  | 0.36  | 0.159     | ns  |
| Budget utilization          | 0.980     | 0.969 | 0.36  | 0.142     | ns  |
| Sentence opener diversity   | 0.973     | 1.00  | -0.31 | n/a       | n/a |
| “the patient” count         | 1.55      | 1.10  | 0.31  | 0.344     | ns  |
| Sentence count              | 4.75      | 4.45  | 0.26  | 0.496     | ns  |
| Adposition ratio            | 10.9%     | 10.1% | 0.25  | 0.576     | ns  |
| Determiner ratio            | 8.5%      | 7.7%  | 0.23  | 0.618     | ns  |
| Certainty markers           | 0.60      | 0.45  | 0.20  | 0.726     | ns  |
| Adverb rate                 | 2.4%      | 2.9%  | -0.19 | 0.576     | ns  |
| Subjectivity                | 0.380     | 0.426 | -0.17 | 0.952     | ns  |
| Epistemic hedging           | 0.45      | 0.35  | 0.13  | 0.952     | ns  |
| Sentiment polarity          | 0.028     | 0.017 | 0.12  | 0.576     | ns  |
| Warning expressions         | 0.85      | 1.10  | -0.12 | 0.969     | ns  |
| Total hedging               | 0.50      | 0.40  | 0.11  | 0.952     | ns  |
| Contrastive connectives     | 0.20      | 0.25  | -0.10 | 1.0       | ns  |
| Verb rate                   | 13.7%     | 13.5% | 0.04  | 0.969     | ns  |
| Total connectives           | 1.45      | 1.40  | 0.03  | 0.867     | ns  |
| Average parse depth         | 5.30      | 5.32  | -0.01 | 1.0       | ns  |
| Hedge-certainty ratio       | 0.458     | 0.458 | 0.00  | 0.969     | ns  |
| Reassurance expressions     | 0.60      | 0.60  | 0.00  | 1.0       | ns  |
| Evidential hedging          | 0.05      | 0.05  | 0.00  | n/a       | n/a |
| Meta-referential exprs      | 0.00      | 0.00  | 0.00  | n/a       | n/a |
| Non-ASCII characters        | 0.00      | 0.00  | 0.00  | n/a       | n/a |
| Second-person pronouns      | 0.00      | 0.00  | 0.00  | n/a       | n/a |

Table 6: All 45 linguistic features for the submitted configuration versus the clinician reference, sorted by Cohen’s  $d$  magnitude.  $p_{adj}$ : Benjamini-Hochberg adjusted  $p$ -values (Wilcoxon signed-rank test). Significance: \*\*\*  $p_{adj} < 0.001$ , \*\*  $p_{adj} < 0.01$ , \*  $p_{adj} < 0.05$ , ns = not significant, n/a = test not applicable.

# tt501 at ArchEHR-QA 2026: Few-Shot Prompting with Retrieval-Augmented Generation for Grounded Clinical EHR Question Answering

Tai Tran Tan

University of Information Technology, Ho Chi Minh City, Vietnam

Vietnam National University, Ho Chi Minh City, Vietnam

22521287@gm.uit.edu.vn

## Abstract

We present the system developed by team tt501 for the ArchEHR-QA 2026 shared task, participating in three subtasks: Evidence Identification (Subtask 2), Answer Generation (Subtask 3), and Evidence Alignment (Subtask 4). All approaches rely exclusively on prompt engineering with reasoning-capable large language models from xAI (*grok-4-fast-reasoning* and *grok-4-1-fast-reasoning*) via API, requiring no task-specific fine-tuning. For Subtask 2, we propose a two-stage hybrid pipeline combining BM25 keyword retrieval with full-context Chain-of-Thought prompting and a secondary LLM refinement step, achieving a best Strict Micro-F1 of 58.84. For Subtask 3, we implement a retrieval-augmented generation (RAG) approach that chains predicted evidence from Subtask 2 with few-shot prompting to produce grounded 75-word clinical answers, reaching an overall score of 31.36. For Subtask 4, we develop a recall-optimised few-shot prompting strategy with reasoning annotations embedded in exemplars, enabling the model to correctly identify many-to-many evidence-answer alignments and achieving a best Micro-F1 of 79.11. Our results demonstrate that carefully engineered prompts with reasoning-capable LLMs can achieve competitive performance on fine-grained clinical evidence tasks without any task-specific training.

**Keywords:** clinical NLP, EHR question answering, evidence identification, retrieval-augmented generation, few-shot prompting, evidence alignment

## 1. Introduction

Responding to patient-generated messages through electronic health record (EHR) portals places a growing burden on clinical staff (Sinsky et al., 2016). A system capable of automatically answering such questions must do more than produce a fluent response: it must locate the relevant clinical evidence within the patient's own notes and present an answer that is directly grounded in that evidence. This requirement makes the problem inherently multi-step, touching on information retrieval, faithful text generation, and fine-grained citation.

The ArchEHR-QA 2026 shared task (Soni and Demner-Fushman, 2026b) decomposes this challenge into four subtasks, extending the 2025 edition that attracted 29 participating teams (Soni et al., 2025). We submitted systems for three of the four subtasks, forming a natural end-to-end pipeline. Subtask 2 asks systems to select the minimal set of note sentences that provide clinical evidence for a given patient question. Subtask 3 uses that evidence to generate a concise, grounded answer. Subtask 4 then requires attributing each sentence in the answer back to the supporting note sentences, producing an explicit citation structure.

Our approach is deliberately lightweight: all three subtasks are addressed through prompt engineering alone, using the *grok-4-1-fast-reasoning* model (xAI, 2025) via the xAI API,

without any fine-tuning or supplementary training data. The central question we investigate is how much can be achieved through careful prompt design when the underlying model already has strong reasoning capabilities. This paper describes our system and discusses what worked and why across each subtask.

## 2. Task and Data

The ArchEHR-QA 2026 dataset (Soni and Demner-Fushman, 2026a) consists of 167 patient cases derived from the MIMIC database (Johnson et al., 2016). Each case includes a patient-authored question, a clinician-interpreted reformulation of that question (at most 15 words), a clinical note excerpt with sentence-level relevance labels (*essential*, *supplementary*, or *not-relevant*), a reference clinician-authored answer, and gold evidence-answer alignment annotations. The dataset is divided into a development set (cases 1–20) and two test splits (cases 21–120 and 121–167).

The shared task defines four subtasks. We participate in the three that form a coherent retrieval-generation-attribution pipeline and do not submit to Subtask 1 (question interpretation). The three subtasks addressed in this work are:

**Subtask 2 (Evidence Identification)** Given the note and the patient question, systems must predict the minimal set of note sentences sufficient to



answer the question. Performance is measured by Precision, Recall, and F1 over predicted sentence-ID sets under two variants: *Strict* (essential sentences only) and *Lenient* (essential and supplementary sentences combined). The official test split covers 47 cases (IDs 121–167).

**Subtask 3 (Answer Generation)** Using the identified evidence, systems must generate a response of at most 75 words that is directly grounded in the note. Answers are evaluated against reference clinician responses using BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), SARI (Xu et al., 2016), BERTScore (Zhang et al., 2020), AlignScore (Zha et al., 2023), and MEDCON (Ben Abacha et al., 2023); the overall score is the arithmetic mean of all six. The evaluation covers the same 47 cases as Subtask 2.

**Subtask 4 (Evidence Alignment)** Given a generated answer and the associated note, systems must align each answer sentence to the note sentence(s) from which it derives, producing a many-to-many mapping. Performance is measured by Micro and Macro Precision, Recall, and F1 over predicted sentence-pair links. The evaluation covers 147 cases (IDs 21–167).

### 3. System Description

All three subtasks use the xAI API with temperature set to 0 for deterministic outputs and JSON-structured responses parsed with regex fallbacks for robustness.<sup>1</sup> Two model versions were used: `grok-4-fast-reasoning` for earlier runs and `grok-4-1-fast-reasoning` (xAI, 2025) for subsequent runs, both reasoning-capable large language models from xAI. For each subtask we submitted two runs, designed so that the second run directly addresses a known limitation of the first. All scripts and prompt templates are publicly available at [urlhttps://github.com/taitran501/tt501-archehr-qa-2026](https://github.com/taitran501/tt501-archehr-qa-2026).

#### 3.1. Subtask 2: Evidence Identification

Selecting the right evidence sentences requires balancing recall against precision: missing a relevant

<sup>1</sup>The xAI API returns JSON reliably at temperature 0; regex fallback was triggered in fewer than 3% of responses, exclusively due to occasional leading/trailing whitespace or markdown code-fence artefacts wrapping an otherwise valid JSON object. The fallback strips these wrappers and re-parses; if parsing still fails the case is logged and excluded rather than silently producing incorrect output.

sentence is just as harmful as including an irrelevant one. We explored two strategies that differ primarily in how much of the note is exposed to the language model.

**Run 1: Hybrid BM25 and LLM Reranker (*hybrid-bm25-llm*).** The first run uses BM25 (Robertson and Zaragoza, 2009) as a first-stage retriever, selecting the top-20 candidate sentences per case using the clinician-interpreted question as the query. These candidates are then passed to `grok-4-fast-reasoning` in a single API call with a precision-focused prompt that instructs the model to return only the sentence IDs that directly and sufficiently answer the question. This approach is computationally efficient, processing 47 cases in approximately three minutes, but it inherits BM25’s sensitivity to vocabulary mismatch: evidence expressed in clinical terminology that does not appear verbatim in the question can be filtered out before the LLM ever sees it.

**Run 2: Full-Context Chain-of-Thought Ensemble with Refinement (*ensemble-refine-v1*).** The second run presents the complete note excerpt to the LLM, removing the first-stage filter entirely. The prompt is oriented toward recall, instructing the model to trace the full clinical narrative (problem, intervention, complication, and resolution) and to include all causally relevant sentences. Two annotated development examples serve as few-shot demonstrations. A secondary LLM call then reviews the initial selection, filtering out noise and resolving cases where patient phrasing and clinical documentation use different terminology for the same concept.

#### 3.2. Subtask 3: Answer Generation

Subtask 3 requires generating a professional, grounded answer of at most 75 words, without introducing clinical claims that are not supported by the note.

**Run 1: Zero-Shot Full-Note Generation (*zero-shot-v1*).** The first run feeds the entire note excerpt to `grok-4-fast-reasoning` without any pre-selected evidence or few-shot examples. The system prompt specifies the 75-word limit, the requirement for a professional medical register, and strict grounding to the note. Output is requested as a JSON object with an `answer` field to prevent formatting artefacts.

**Run 2: Retrieval-Augmented Generation with Few-Shot Examples (*rag-fewshot-v1*).** The second run implements a retrieval-augmented generation pipeline that chains directly from the Subtask 2

output. The sentences predicted as evidence by *ensemble-refine-v1* are resolved to their text and presented as a structured block, reducing the input to only the sentences considered relevant. Three gold-annotated development cases are included as few-shot demonstrations, each pairing the patient question, the evidence sentences, and the reference clinician answer. This design encourages the model to produce responses that are lexically and semantically close to how a clinician would write, and the reduced context appears to help calibrate response length.

### 3.3. Subtask 4: Evidence Alignment

Evidence alignment requires mapping each answer sentence to the note sentence(s) from which it was derived. The task is many-to-many: a single answer sentence may synthesise information from several parts of the note, and under-citing is a common failure mode when a prompt only rewards identifying the most lexically similar match.

**Run 1: Basic Few-Shot Alignment (*grok-4-1-reasoning-fewshot-v1*).** The first run supplies the clinician-interpreted question, the note excerpt, and the answer sentences to the model. A few-shot prompt with development examples illustrates the expected JSON output format consisting of `{answer_id, evidence_id}` objects. This provides a clean baseline for how much the model can align without explicit guidance on *how* to reason about citations.

**Run 2: Recall-Optimised Few-Shot with Reasoning Annotations (*few-shot-reasoning-v2*).** The second run introduces two changes over Run 1. First, the patient-authored question is added alongside the clinician question, because the patient’s own wording often mirrors the lay language used in the answer and can serve as a bridge to clinical terminology in the note. Second, and more importantly, each of the four few-shot examples is augmented with explicit reasoning: for each citation link, the prompt explains in plain language why that note sentence supports that answer sentence. This teaches the model the rationale for citing, not only the format. The system prompt also includes an explicit instruction that answer sentences frequently synthesise information from multiple note sentences, counteracting the natural tendency to select only the most lexically similar source.

## 4. Results and Analysis

Tables 1–3 summarise official test-set scores for each submitted run. Run 1 of Subtasks 2 and 3 uses *grok-4-fast-reasoning*; all other runs

| Run                | Strict |      |             | Lenient |      |             |
|--------------------|--------|------|-------------|---------|------|-------------|
|                    | P      | R    | F1          | P       | R    | F1          |
| hybrid-bm25-llm    | 50.1   | 56.0 | 52.9        | 66.7    | 56.0 | 60.9        |
| ensemble-refine-v1 | 56.0   | 61.9 | <b>58.8</b> | 74.9    | 61.9 | <b>67.8</b> |

Table 1: Subtask 2 test results (IDs 121–167,  $n=47$ ). Micro Precision (P), Recall (R), F1 (%).

| Run            | BLEU | R-L  | SARI | BERT | MED  | ALN  | Ovrl        |
|----------------|------|------|------|------|------|------|-------------|
| zero-shot-v1   | 3.0  | 19.0 | 53.5 | 36.8 | 36.6 | 19.4 | 28.1        |
| rag-fewshot-v1 | 6.2  | 24.0 | 58.3 | 41.4 | 35.9 | 22.4 | <b>31.4</b> |

Table 2: Subtask 3 test results ( $n=47$ ). All scores in %. BERT=BERTScore, MED=MEDCON, ALN=AlignScore, Ovrl=overall (mean of all six).

use *grok-4-1-fast-reasoning*. All experiments run with temperature 0 and no external resources beyond the shared-task data.

### 4.1. Subtask 2: Evidence Identification

The Subtask 2 results highlight the value of exposing the model to the full clinical note rather than a BM25-filtered subset. With complete context and a Chain-of-Thought prompt, the ensemble run (Run 2) achieves a Strict Micro-F1 of 58.8 and a Lenient F1 of 67.8 (Table 1), indicating that the model can reliably identify sentences that carry the core clinical evidence. Its Strict recall of 61.9 suggests that many relevant sentences are recovered even when they are only indirectly related to the surface wording of the question, a setting in which lexical retrieval alone is brittle. The consistent gap between Lenient and Strict F1 across runs further points to the presence of sentences that provide background or clarifying information: they are not strictly necessary for correctness, but they enrich the clinical story presented to the patient.

We note that the two runs also differ in model version: Run 1 uses *grok-4-fast-reasoning*, whereas Run 2 uses *grok-4-1-fast-reasoning*. The observed improvement in Strict F1 from 52.9 to 58.8 therefore reflects the combined effect of the switch to full-context prompting *and* the newer model version, and cannot be attributed exclusively to prompt design. Isolating the contribution of each factor would require rerunning both prompts under the same model version, which we leave for future work.

### 4.2. Subtask 3: Answer Generation

In Subtask 3, the contrast between the two runs reflects the benefit of separating retrieval from generation and showing the model concrete clinical examples. When the model receives only the full note and a zero-shot prompt, it must jointly decide which parts of the note matter and how to phrase

| Run        | Micro |      |             | Macro |      |             |
|------------|-------|------|-------------|-------|------|-------------|
|            | P     | R    | F1          | P     | R    | F1          |
| fewshot-v1 | 85.1  | 72.5 | 78.3        | 86.4  | 75.6 | 78.5        |
| fewshot-v2 | 78.2  | 80.1 | <b>79.1</b> | 80.0  | 82.6 | <b>79.7</b> |

Table 3: Subtask 4 test results (IDs 21–167,  $n=147$ ). Precision (P), Recall (R), F1 (%). *fewshot-v1*: basic few-shot, clinician question only. *fewshot-v2*: recall-optimised, reasoning annotations, patient+clinician question.

the answer. Conditioning instead on sentences already identified as evidence, together with three few-shot demonstrations, leads to answers that better align with the reference along lexical, semantic, and factual dimensions: the RAG few-shot run (Run 2) attains an overall score of 31.4 compared with 28.1 for the zero-shot baseline, with consistent improvements across BLEU, ROUGE-L, SARI, BERTScore, MEDCON, and AlignScore (Table 2). The behaviour around the 75-word constraint reinforces this picture: 16 zero-shot predictions require truncation, whereas only a single RAG prediction does so, suggesting that the demonstrations also help the model internalise an appropriate answer length and professional register.

As with Subtask 2, the two runs differ in model version: Run 1 uses `grok-4-fast-reasoning` and Run 2 uses `grok-4-l-fast-reasoning`. Consequently, the gain in overall score from 28.1 to 31.4 reflects both the switch to a RAG pipeline with few-shot examples and the use of an updated model, and neither factor can be fully isolated from the other.

### 4.3. Subtask 4: Evidence Alignment

For evidence alignment (Table 3), the recall-optimised prompt with reasoning annotations encourages the model to treat citation as a genuinely many-to-many mapping rather than a one-to-one matching problem. The corresponding run (Run 2) reaches a Micro-F1 of 79.1, with recall of 80.1 and precision of 78.2, while the basic few-shot baseline (Run 1) attains a Micro-F1 of 78.3 with recall and precision of 72.5 and 85.1. This pattern is consistent with the prompt’s instruction to cite *all* sentences that contribute to a claim: some precision is traded for recall, but the overall balance better reflects the evaluation protocol, which penalises both missing and spurious links. The reasoning annotations in the exemplars make this behaviour concrete by spelling out how each citation connects an answer span to specific parts of the note, and the model appears to emulate this strategy at test time. Adding the patient-authored question to the context further helps bridge lay phrasing in the answer and clinical terminology in the note, making

it easier for the model to recognise when multiple note sentences jointly support a single statement.

### 4.4. Qualitative Error Analysis

To clarify where each approach succeeds and remains fragile, we examine representative failure cases drawn from the development set.

#### Subtask 2: vocabulary mismatch in BM25 retrieval.

The hybrid run (Run 1) uses BM25 with the clinician question as query, filtering the note excerpt to a top-20 candidate set before the LLM sees any evidence. In Case 4 (cardiac catheterisation), the clinician question asks “*Why was cardiac catheterization recommended?*” One gold-essential sentence reads: “*Last echo showed LVEF = 25%*”, the key clinical finding that motivated the procedure, but the abbreviations *LVEF* and *echo* do not appear in the question. BM25 assigns it a low retrieval score, and the sentence is below the top-20 cut-off, so the LLM reranker never sees it. The ensemble run (Run 2) presents the full note, and the CoT prompt’s instruction to trace the complete clinical narrative (“*What led to this decision?*”) correctly recovers the sentence. This pattern (essential context buried in clinical abbreviations not shared with the patient-facing question) recurs across cardiovascular and ICU cases and accounts for a substantial fraction of Run 1’s false negatives.

#### Subtask 3: incomplete grounding in zero-shot generation.

The zero-shot run (Run 1) is prone to generating answers that cover only the first relevant thread in the evidence, ignoring subsequent instructions or alternative findings. In Case 3 (traumatic brain injury discharge instructions), the gold answer spans five sentences: expected symptoms (drowsiness, headaches, dizziness, irritability, memory loss), a prognosis note (symptom decrease over weeks), a follow-up instruction (TBI specialist referral), a general recovery reminder, and a list of warning signs requiring urgent re-contact (visual changes, unilateral weakness, speech difficulty). Run 1 produced a single sentence: “*Symptoms such as irritability and headaches are normal after a traumatic brain injury and should decrease over the next several weeks*” – correct but radically incomplete. The RAG few-shot run (Run 2) mitigates this by conditioning on pre-selected evidence sentences and providing three complete reference answers as exemplars, which anchor the model’s expected output length and content breadth.

#### Subtask 4: under-citation of synthesised answer sentences.

The basic few-shot baseline (Run 1) confidently assigns a single note sentence to each answer sentence, producing high precision

(85.1) at the cost of recall (72.5). In Case 5 (musculoskeletal chest pain), the first answer sentence *“The chest pain was musculoskeletal, not overdose or cardiac related”* is a conclusion synthesised from four separate note sentences: the musculoskeletal exam finding, the normal EKG, the negative cardiac enzymes, and the normal transthoracic echocardiogram. Run 1 cites only the exam finding sentence ([9]). The recall-optimised run (Run 2) includes reasoning annotations in its exemplars that explicitly model this synthesis pattern (e.g., *“Answer mentions no cardiac causes: cite EKG sentence [10], enzyme sentence [18], and echo sentence [19]”*), teaching the model that a conclusory answer sentence can require multi-source support, and thereby recovering the missing citations.

#### 4.5. Inference Cost and Latency

Both `grok-4-fast-reasoning` and `grok-4-1-fast-reasoning` share the same pricing tier on the xAI API: \$0.20 per million input tokens and \$0.50 per million output tokens (reasoning tokens billed at the output rate). Submitting via the Batch API reduces all token costs by 50% relative to synchronous calls.

Table 4 summarises the API usage pattern for each submitted run. Subtask 2 Run 1 processes all 47 cases via synchronous API calls with a BM25 pre-filter, completing in approximately three minutes. All other runs use the xAI async Batch API, submitting all cases in a single batch job and polling for results, benefiting from the 50% batch discount. Batch jobs for the 47-case Subtasks 2–3 sets typically completed within approximately 10 minutes. The 147-case Subtask 4 batches required approximately 15 minutes, reflecting the larger per-case input context. Input contexts for the full-context runs (Subtask 2 Run 2 and both Subtask 3 runs) are substantially larger than those of the BM25-filtered run because the entire note excerpt rather than a top-20 candidate subset is provided to the model.

Despite the reasoning model overhead and the full-context prompts, the combined cost of each individual run remained below \$0.20 USD, and the total spend across all six test-set runs was under \$1.00. This demonstrates that prompt-engineering approaches on small shared-task datasets are viable even on a personal API budget.

### 5. Conclusion

This paper described a prompt-only system for three subtasks of the ArchEHR-QA 2026 shared task. Working entirely through prompt engineering on a single reasoning-capable language model, without any task-specific fine-tuning, we found that

| Subtask | Run                | Cases | API Mode    | Time    | Cost    |
|---------|--------------------|-------|-------------|---------|---------|
| 2       | hybrid-bm25-llm    | 47    | Synchronous | ~3 min  | <\$0.20 |
| 2       | ensemble-refine-v1 | 47    | Async Batch | ~10 min | <\$0.20 |
| 3       | zero-shot-v1       | 47    | Async Batch | ~10 min | <\$0.20 |
| 3       | rag-fewshot-v1     | 47    | Async Batch | ~10 min | <\$0.20 |
| 4       | fewshot-v1         | 147   | Async Batch | ~15 min | <\$0.20 |
| 4       | fewshot-v2         | 147   | Async Batch | ~15 min | <\$0.20 |

Table 4: Inference cost and latency per run on the official test sets. Wall-clock time is approximate and subject to xAI server load. Cost is based on the xAI pricing at competition time: \$0.20/\$0.50 per million input/output tokens (standard); Batch API runs benefit from a 50% discount. Total spend across all six runs was under \$1.00.

the design of the prompt itself has a substantial effect on performance in each subtask.

For evidence identification, the key insight is that restricting the model’s input to BM25-selected candidates limits recall in ways that cannot be recovered by downstream reranking. Allowing the model to read the full note and reason over the complete clinical narrative is more effective, even at a higher inference cost. For answer generation, few-shot demonstrations drawn from annotated development cases help the model adopt the register and length expected by clinical evaluators, and coupling generation with retrieved evidence produces answers that are more faithful to the source note. For evidence alignment, the most impactful change was adding explicit reasoning to the few-shot exemplars: rather than showing the model what correct citations look like, explaining why each citation is correct guides the model to reason about multi-source attribution rather than defaulting to surface lexical similarity.

Taken together, these results support the broader observation that reasoning-capable language models can be effective on structured clinical NLP tasks when prompts are designed to make the reasoning process explicit, rather than leaving the model to infer it from output examples alone. Future work will examine joint optimisation across the subtask pipeline, self-consistency strategies for evidence identification, and principled methods for selecting few-shot demonstrations.

### 6. Limitations

**Reproducibility.** All results depend on the `grok-4-1-fast-reasoning` and `grok-4-fast-reasoning` APIs provided by xAI. Model outputs may change as providers update or deprecate model versions, reducing exact reproducibility over time. Temperature is set to 0 throughout to mitigate non-determinism, but API-level changes remain outside our control.



**Model version confound.** Run 1 of Subtasks 2 and 3 uses `grok-4-fast-reasoning`, whereas all other runs use `grok-4-1-fast-reasoning`. The observed improvements in Subtask 2 (Strict F1: 52.9 → 58.8) and Subtask 3 (overall: 28.1 → 31.4) therefore reflect the combined effect of prompt and pipeline changes *and* a model version upgrade; the contribution of each factor cannot be fully isolated without rerunning all configurations under the same model, which we leave for future work. The mixed versioning was intentional: `grok-4-fast-reasoning` was the model available when Subtasks 2 and 3 development began. As experimentation progressed it was superseded by `grok-4-1-fast-reasoning`, which xAI positioned as the recommended reasoning successor; Subtask 4 was designed from the outset to use `grok-4-1-fast-reasoning` for both runs, so performance differences there reflect only prompt and pipeline changes. As of March 2026, `grok-4-fast-reasoning` is no longer listed on the active xAI models page; researchers replicating this work should use `grok-4-1-fast-reasoning` or its current successor.

**Error propagation in the RAG pipeline.** Subtask 3 results are conditioned on Subtask 2 predictions. Errors in evidence identification directly degrade the quality of generated answers. An analysis of cases where evidence sets were incomplete or incorrect is left for future work.

**No human evaluation.** Our system is assessed solely through automatic metrics. For clinical applications, human evaluation by domain experts would be necessary to verify factual accuracy, appropriate professional register, and patient safety. This limitation is particularly consequential for Subtask 3, where the generated answer is presented directly to a patient. Automatic metrics such as BLEU, ROUGE-L, and BERTScore measure lexical and semantic overlap with a reference answer but cannot detect clinically unsafe content, for example an omitted contraindication, a misattributed finding, or phrasing that could be misinterpreted in a self-care context. Human review by a clinician expert would be required before any such system could be considered for deployment in a patient-facing setting.

**Dataset size.** The ArchEHR-QA 2026 test set contains 47 cases for Subtasks 2–3 and 147 cases for Subtask 4. Score differences of <1 point should be interpreted with caution given the limited sample sizes.

## 7. Ethical Considerations

The ArchEHR-QA dataset is a de-identified derivative of the MIMIC database (Johnson et al., 2016), distributed by the shared-task organisers for competition use. All clinical data were de-identified prior to inclusion in the released corpus; no fields that could re-identify individuals are present.

During experimentation, this data was submitted to the xAI API under an individual account. We note that the xAI consumer Terms of Service permit use of inputs for model improvement by default, and the PhysioNet Data Use Agreement recommends verifying zero-retention guarantees before transmitting MIMIC-derived data to any third-party service. We acknowledge this as an unresolved compliance concern. Researchers replicating this work are strongly advised to use an enterprise API agreement with explicit zero-retention and no-training guarantees, or a locally hosted model, and to independently verify alignment with the PhysioNet DUA applicable to their own credentials.

This work does not introduce new patient data, does not perform any re-identification attempts, and does not deploy the system in a clinical setting. The system is intended solely for research purposes and should not be used for clinical decision-making without further validation.

## 8. Bibliographical References

- Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An investigation of evaluation metrics for automated medical note generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 15128–15143.
- 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.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.
- 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 (ACL 2002)*, pages 311–318.



Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760.

Sarvesh Soni and Dina Demner-Fushman. 2026a. [A dataset for addressing patient’s information needs related to clinical course of hospitalization](#). *Scientific Data*.

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)*. ELRA.

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 BioNLP Workshop at ACL 2025*.

xAI. 2025. Grok-4: Advancing reasoning and scientific discovery. <https://x.ai/blog/grok-4>. Accessed 2026.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 401–415.

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 (ACL 2023)*, pages 11328–11348.

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 (ICLR)*.

## A. Prompts and Few-Shot Examples

This appendix provides the exact system prompts, user prompt templates, and few-shot examples used for each submitted run. Variable placeholders are shown in `{curly_braces}`. All runs use `temperature=0` and JSON-structured output.

### Subtask 2, Run 1: *hybrid-bm25-llm*

The BM25 first stage selects the top-20 candidates using the clinician question as the query. The LLM reranker then receives the following prompt, where `{clinician_question}` is the question text and the note is limited to the BM25-selected sentences.

#### System message.

*You are a clinical documentation specialist.*

#### User message.

**Question:** `{clinician_question}`

**Task:** Identify all sentences in the Note Excerpt below that provide evidence answering this question.

**Output Format:** Respond ONLY with a JSON object: `{"evidence_ids": ["0", "2", ...]}`

### Subtask 2, Run 2: *ensemble-refine-v1*

The full note excerpt is provided to the model. The prompt is recall-oriented and uses two annotated development examples as few-shot demonstrations.

#### System message.

*You are a precise medical evidence classifier. Respond ONLY with valid JSON.*

#### User message template.

You are an expert medical documentation specialist. Your task is to identify the COMPREHENSIVE set of sentences from the clinical note that provide EVIDENCE to answer the clinician’s question.

##### Instructions:

1. Read the **Clinician Question** and the **Clinical Note** carefully.
2. Select **ALL** sentences that: (a) provide direct answers, (b) provide necessary context (timeline, initial conditions, reasons for changes), (c) explain WHY a decision was made or HOW an outcome occurred, (d) detail specific values, dates, or test results relevant to the inquiry.
3. Completeness is critical. It is better to include a borderline sentence than to miss it.
4. Capture the full narrative arc: Problem → Intervention → Complication → Resolution.

### Example 1

Note: [1] During the ERCP a pancreatic stent was required. . . [7] As the patient's INR was normalized to 1.2, a sphincterotomy was safely performed, with removal of several biliary stones. . .

```
Answer: {"relevant_ids": ["1",
"5", "6", "7"]}
```

### Example 2

Note: [1] Acute diastolic heart failure: Pt developed signs and symptoms of volume overload... [6] This undoubtedly contributed to his continued hypoxemic respiratory failure.

**Answer:** {"relevant\_ids": ["1", "4", "5", "6"]}

### Current Task:

**Clinician Question:** {question}

**Clinical Note:**

```
{note text}
```

### Analyze step-by-step:

1. Identify the core inquiry.
2. Trace the relevant storyline (chronological, cause-and-effect).
3. Select ALL sentence IDs that form the complete evidence chain.

**Output Format:** Respond ONLY with valid JSON: {"relevant\_ids": ["1", "3", "5"]}

The entire note excerpt is provided; no pre-selected evidence or few-shot examples are used.

*You are a clinical documentation specialist extracting factual answers from medical notes.*

Critical Guidelines:

1. Extract **ONLY** information explicitly stated in the note—no speculation, inference, or hedging.
2. Use professional medical register (“The patient...”). No empathy phrases.
3. 75 words maximum (strictly enforced).
4. Use past tense for completed medical events.
5. State facts directly without qualifiers.

**Patient Question:** {patient question}

**Clinician Question:** {clinician\_question}

**Note Excerpt:** {full\_note\_text}

**Output Format:** Respond ONLY with valid

```
JSON: {"answer": "Your factual
extraction here (max 60 words)"}
```

Evidence sentences predicted by *ensemble-refine-v1* are resolved to their text and passed instead of the full note. Three development-set reference answers serve as few-shot demonstrations.

*You are a professional medical scribe. Your task is to draft a concise clinical answer to a clinician's question, based ONLY on the provided evidence sentences.*

Constraints:

1. **Length:** Maximum 75 words ( $\approx$ 5 sentences).
2. **Tone:** Professional, detached, medical register ("The patient. . ."). Do NOT use "You" or conversational language.
3. **Grounding:** Use ONLY the provided evidence. Do not hallucinate or use external knowledge.
4. **Format:** Plain text paragraph. No citations. No markdown.

**User message template (few-shot prefix).**

Three examples taken from the development set are prepended (abbreviated here for space). The full examples correspond to Cases 4, 1, and 9 from the ArchEHR-QA 2026 development data.

### Example Case:

**Question:** Why was cardiac catheterization recommended to the patient?

**Evidence:**

[selected essential sentences from Case 4]

**Answer:** The patient was recommended a cardiac catheterization for worsening heart failure confirmed by left ventricle ejection fraction of 25% on his echocardiogram...

⟨two further examples omitted for brevity⟩

**Current Case:**

**Question:** {clinician question}

**Evidence:**

```
{evidence text}
```

**Answer:**

Two development-set examples are used. The patient question is omitted from context. No reasoning annotations are included in the exemplars.

## System message.

*You are an expert clinical evidence aligner. Your task is to align each answer sentence with the supporting sentence(s) from the clinical note excerpt. For each answer sentence (identified by Answer ID), return a JSON object containing the Answer ID and an array of Evidence IDs from the Note Sentences that support it. If an answer sentence is not supported by any note sentence, return an empty array [] for its evidence\_id.*

Constraints:

1. Return ONLY valid JSON as a list of objects. Do NOT wrap in Markdown.
2. Format: 

```
[{"answer_id": "1", "evidence_id": ["2", "5"]}, ...]
```
3. Avoid over-citing.
4. Alignments are many-to-many.

## User message template.

---

### Example Case:

**Clinician Question:** {clinician\_question}

**Note Excerpt:**

[1]... [N]...

**Answer Sentences:**

[Answer 1]...

**Output:** [ {"answer\_id": "1", "evidence\_id": ["2"]}, ... ]

(second example follows the same pattern)

---

### Current Case:

**Clinician Question:** {clinician\_question}

**Note Excerpt:**

{note\_text}

**Answer Sentences:**

{answer\_text}

**Output:**

## Subtask 4, Run 2: *fewshot-v2*

Four development-set examples (Cases 1, 2, 5, 6) are used. The patient question is added to context. Exemplars include reasoning annotations explaining why each note sentence supports the corresponding answer sentence.

## System message.

*You are an expert clinical evidence aligner specialised in grounding answer sentences to clinical note excerpts.*

Key Alignment Principles:

1. An answer sentence often synthesises information from MULTIPLE note sentences. When this happens, cite ALL contributing note sentences.
2. A note sentence supports an answer sentence if it provides specific clinical facts referenced in the answer—even if the wording differs (paraphrasing is common).
3. If an answer sentence is a general conclusion NOT grounded in any specific note sentence, assign an empty array [].
4. Do NOT over-cite: only cite note sentences whose information is actually referenced in the answer sentence.

## Output Rules:

1. Return ONLY valid JSON. NO Markdown code fences.
2. Format: 

```
[{"answer_id": "1", "evidence_id": ["2", "5"]}, ...]
```
3. Include an entry for EVERY answer sentence in order.
4. Alignments are many-to-many.

## User message template (abbreviated few-shot prefix).

---

### Example Case (ID 1):

**Patient Question:** ...

**Clinician Question:** Why was ERCP recommended...

**Note Excerpt:**

[1]... [N]...

**Answer Sentences:**

[Answer 1]...

**Reasoning:**

Answer 1: "ERCP was recommended to place a CBD stent" → Note [2] describes ERCP placing a CBD stent.

...

**Output:** [ {"answer\_id": "1", "evidence\_id": ["2"]}, ... ]

(Examples for Cases 2, 5, and 6 follow the same pattern.)

---

### Current Case:

**Patient Question:** {patient\_question}

**Clinician Question:** {clinician\_question}

**Note Excerpt:**

{note\_text}

**Answer Sentences:**

{answer\_text}

**Output:**

# HealthNLP\_Retrievers at ArchEHR-QA 2026: Cascaded LLM Pipeline for Grounded Clinical Question Answering

Md Biplob Hosen<sup>1,2</sup>, Md Alomgeer Hussein<sup>1</sup>, Md Akmol Masud<sup>2</sup>,  
Omar Faruque<sup>1</sup>, Tera L Reynolds<sup>1</sup>, Lujie Karen Chen<sup>1</sup>

<sup>1</sup>University of Maryland Baltimore County, Baltimore, MD, USA

<sup>2</sup>Jahangirnagar University, Savar, Dhaka, Bangladesh

{mhosen1, mdalomh1, omarf1, reynoter, lujiec}@umbc.edu, masud.stu2018@juniv.edu

## Abstract

Patient portals now give individuals direct access to their electronic health records (EHRs), yet access alone does not ensure patients understand or act on the complex clinical information contained in these records. The ArchEHR-QA 2026 shared task addresses this challenge by focusing on grounded question answering over EHRs, and this paper presents the system developed by the HealthNLP\_Retrievers team for this task. The proposed approach uses a multi-stage cascaded pipeline powered by the Gemini 2.5 Pro large language model to interpret patient-authored questions and retrieve relevant evidence from lengthy clinical notes. Our architecture comprises four integrated modules: (1) a few-shot query reformulation unit which summarizes verbose patient queries; (2) a heuristic-based evidence scorer which ranks clinical sentences to prioritize recall; (3) a grounded response generator which synthesizes professional-caliber answers restricted strictly to identified evidence; and (4) a high-precision many-to-many alignment framework which links generated answers to supporting clinical sentences. This cascaded approach achieved competitive results. Across the individual tracks, the system ranked 1st in question interpretation, 5th in answer generation, 7th in evidence identification, and 9th in answer-evidence alignment. These results show that integrating large language models within a structured multi-stage pipeline improves grounding, precision, and the professional quality of patient-oriented health communication. To support reproducibility, our source code is publicly available in our [GitHub repository](#).

**Keywords:** Clinical Question Answering, Electronic Health Records, Large Language Models, Clinical NLP, Evidence Grounding, Cascaded Pipeline.

## 1. Introduction

Large language models (LLMs) have significantly advanced consumer health question answering. Systems such as Med-PaLM 2 achieve expert-level performance on medical licensing benchmarks (Singhal et al., 2025). Despite these advances, general-purpose LLMs often struggle to provide grounded answers. Grounding refers to the ability to anchor responses in a specific patient's medical history without introducing unjustified clinical information (Wang et al., 2025). This is a major limitation in EHRs, demonstrating a clear vocabulary gap. Patient inquiries are often verbose, informal, and emotionally expressive. The clinical notes remain dense, fragmented, and written in a specialized technical language (Nie et al., 2014; Abacha and Demner-Fushman, 2019b). As a result, translating patient-authored questions into clinically meaningful information remains a central challenge for patient-facing health technologies.

The ArchEHR-QA shared task series addresses the communication gap between patient language and clinical documentation in both directions. The task builds on a curated dataset derived from the MIMIC-III critical care database (Johnson et al., 2016; Soni and Demner-Fushman, 2026a). Unlike earlier clinical question answering tasks focused on extracting facts from large medical corpora (Jin

et al., 2019), ArchEHR-QA requires systems to operate directly on patient-specific EHR excerpts. Participating systems must interpret subjective patient narratives, identify minimal supporting evidence within complex clinical documentation, and generate tightly constrained grounded answers (Soni and Demner-Fushman, 2026b).

The initial 2025 iteration attracted 29 participating teams submitting 75 systems (Soni et al., 2025). This participation reflects strong research interest in evidence-grounded clinical question answering. The 2026 iteration expands the challenge to four complementary subtasks: question interpretation, evidence identification, answer generation, and answer-evidence alignment. Models such as the Gemini series demonstrate strong long-context reasoning capabilities within clinical text (Team et al., 2024). Precise alignment between generated answers and sentence-level evidence remains a major challenge (Gao et al., 2023). Research increasingly supports multi-stage architectures which decompose complex reasoning tasks into structured intermediate steps instead of relying on single-pass generation (Ghafoor et al., 2025). Role-based prompting and few-shot demonstrations further guide model outputs toward a professional clinical communication style suitable for patient-facing applications (Singhal et al., 2023; Nori et al., 2023).

This paper presents the HealthNLP\_Retrievers system, an applied engineering contribution to the ArchEHR-QA 2026 shared task. The system employs a structured multi-stage architecture for grounded clinical question answering over EHRs. Our primary contributions are three integrated design choices: (1) a persona-driven interpretive query reformulation step that runs upstream of retrieval; (2) a Likert-scale, recall-biased evidence scoring heuristic with dynamic fallback tiers to prevent downstream evidence starvation; and (3) a precision-constrained many-to-many answer-evidence alignment framework. Moving beyond standard Retrieval-Augmented Generation (RAG) paradigms (Lewis et al., 2020), the system deploys Gemini 2.5 Pro (Comanici et al., 2025) within this four-stage cascaded pipeline to bridge the semantic divide between patient information needs and the objective clinical realities documented in EHRs.

## 2. Related Work

### 2.1. Clinical Question Answering

Automated clinical QA has evolved from structured knowledge-base lookups to open-domain systems capable of reasoning over unstructured text. Early work on consumer health QA focused on summarizing patient questions to bridge the vocabulary gap between nonexpert users and medical professionals (Abacha and Demner-Fushman, 2019b,a). PubMedQA (Jin et al., 2019) introduced a benchmark for biomedical research QA, while more recent efforts have shifted toward patient-specific QA grounded in EHR data (Soni and Demner-Fushman, 2026a). Unlike general health QA, EHR-grounded QA demands that answers are explicitly linked to verifiable clinical evidence, posing unique challenges in retrieval precision and answer fidelity.

### 2.2. LLMs for Medical Reasoning

Large language models have rapidly advanced medical reasoning capabilities. Med-PaLM 2 (Singhal et al., 2025) established expert-level performance on US medical licensing examination style questions, while GPT-4 demonstrated strong diagnostic reasoning across diverse clinical scenarios (Nori et al., 2023). Chain-of-thought (CoT) prompting (Wei et al., 2022) has proven effective for complex clinical reasoning, enabling models to decompose multistep diagnostic logic. Domain-adapted models such as BioBERT (Lee et al., 2020) have shown that pretraining on biomedical corpora significantly improves performance on clinical NLP tasks. More recently, the Gemini model family (Team et al., 2024; Comanici et al., 2025) has demonstrated strong long-context reasoning suited to processing full clinical notes.

### 2.3. Evidence Grounding and RAG

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has emerged as a key paradigm for grounding LLM outputs in external knowledge, reducing hallucinations (Gao et al., 2023). In clinical settings, iterative RAG frameworks such as i-MedRAG (Xiong et al., 2024) allow models to refine queries across multiple retrieval rounds, improving performance on complex medical questions. Standard RAG pipelines retrieve information at the document or passage level, whereas clinical evidence grounding requires sentence-level precision, a gap our Likert-scale scoring approach directly addresses.

### 2.4. Prior ArchEHR-QA Systems

The ArchEHR-QA 2025 shared task (Soni et al., 2025) revealed several effective design patterns among the 29 participating teams. LAMAR (Yoadsanit et al., 2025) employed clinically aligned few-shot learning to generate grounded responses from EHRs, while CUNI (Lanz and Pecina, 2025) investigated whether smaller, more efficient LLMs matched the performance of frontier models on clinical QA. These studies collectively demonstrated that task decomposition and careful prompt engineering often prove more impactful than model scale alone, an insight that informed our cascaded pipeline design.

## 3. Methodology

The proposed HealthNLP\_Retrievers system is a four-stage cascaded pipeline designed to bridge the semantic gap between subjective patient narratives and objective clinical documentation. The system is built using the Gemini 2.5 Pro model to utilize its extended context window to process significantly larger clinical notes. Our approach emphasizes interpretative grounding, interpreting the user's intent ahead of retrieving evidence, and generates a heavily constrained answer. The complete architecture and workflow of this cascaded pipeline are illustrated in Figure 1.

### 3.1. Few-shot Question Interpretation

Patient-authored questions are frequently characterized by high verbosity and informal phrasing, which introduce significant noise into downstream retrieval tasks, while emotional context often provides important signals about patient concerns and needs (Abacha and Demner-Fushman, 2019b). To mitigate retrieval noise while preserving clinically meaningful context, we implemented a query interpretation module.



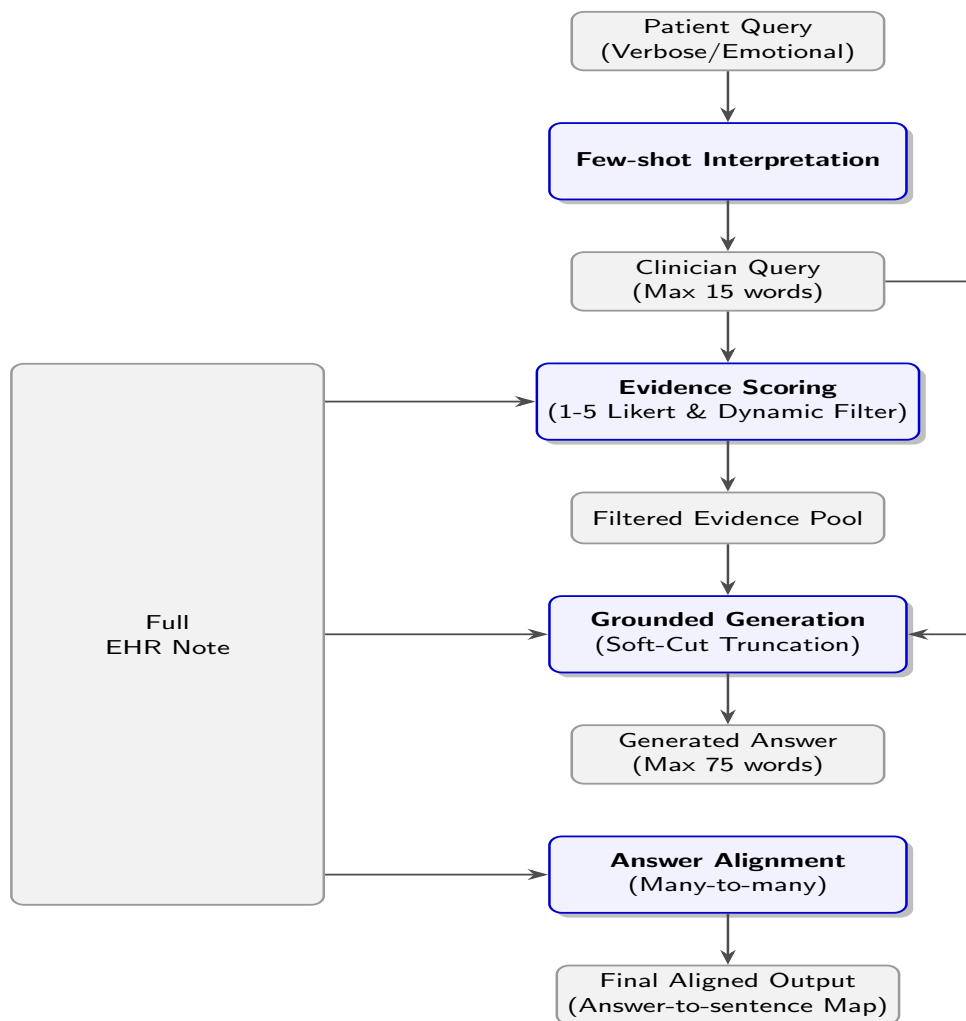


Figure 1: Workflow of the HealthNLP\_Retrievers multi-stage cascaded pipeline.

We employ a clinical administrative assistant persona to transform raw narratives into concise, professional queries. The module utilizes a few-shot prompting strategy, constructing a prompt with three examples sampled from the training set. The prompt enforces strict rules including no preambles, standardized usage of pronouns, and a prohibition against diagnostic language, following structured prompting practices used in prior LLM prompting frameworks. To guarantee compliance with the shared task constraints, a post-processing step strictly truncates the generated output to a maximum of 15 words.

### 3.2. Heuristic-based Evidence Scoring

Standard boolean retrieval is insufficient for clinical QA, where evidence is scattered across disjointed note segments. We reformulated evidence identification as a sentence-level scoring task. We instructed the model to act as a clinical evidence scorer and assign a relevance score (1–5) to ev-

ery sentence, utilizing few-shot examples built from gold-standard annotations. Specifically, the model assigns a score of 5 (critical) to sentences containing direct answers, confirmed diagnoses, or specific medication outcomes. A score of 4 (high context) is allocated to sentences qualifying a critical finding, such as test result context. A score of 3 (relevant) captures background context, admission dates, or clinical history. Finally, scores of 1 and 2 (irrelevant/weak) are reserved for normal findings, administrative headers, or unrelated systems.

**Dynamic Recall-biased Filtering:** To address the high cost of false negatives in healthcare (Burt et al., 2017), we implemented a dynamic filtering mechanism. The system first isolates strict evidence (scores  $\geq 4$ ). If no strict evidence is found, it falls back to a lenient set (scores  $\geq 3$ ) to prevent downstream starvation. In the event of an API blockage or formatting error, the system relies on a simple fallback: selecting the first three sentences of the note.

### 3.3. Grounded Answer Generation

The generation module drafts a concise, professional answer capped at 75 words. By prompting the model with a clinical documentation specialist persona, we ensured the final output maintained a strictly objective tone. Specifically, this persona functions as a zero-shot constraint that restricts the model to precise medical terminology, eliminates conversational preambles, and strictly grounds the output to prevent the inclusion of external medical knowledge.

**Retrieval-Augmented Generation (RAG):** The prompt is constructed with three distinct context blocks: (1) the clinical interpretation of the patient question, (2) the explicitly filtered evidence text from stage 2, and (3) the full clinical note as a secondary context. We explicitly instruct the model to prioritize the filtered evidence while the full note serves as a safety net for resolving implicit references—a deliberate "focus-then-expand" design that reduces hallucinations while preventing under-specification. This dual-context design is consistent with the official task input specification (Soni and Demner-Fushman, 2026b).

**Soft-cut Truncation Heuristic:** Given the strict 75-word constraint, we implemented a custom truncation logic. Post-generation, if the text exceeds the limit, it is hard-cut at 75 words. We prefer to truncate the text cleanly at the end of a sentence. The script looks for the last available period, and if cutting there retains at least 60% of the maximum length, we accept the shorter version. If not, we enforce the hard limit and manually append a period so the output doesn't look broken.

### 3.4. Answer-Evidence Alignment

In the final stage, we tackle many-to-many alignment between the answers and the source text. For the official Subtask 4 evaluation (Soni and Demner-Fushman, 2026b), the system takes pre-segmented clinician answers and the full clinical note as inputs. While a real-world deployment would feed the output of Subtask 3 directly into this stage, we adhered strictly to the shared-task format for testing. We prompted the model to link evidence conservatively, ensuring it only cites sentences that directly back up the answer without pulling in unnecessary background details. To demonstrate this exactness, we embedded a static set of few-shot examples from the development data. The system returns a JSON array that maps every answer sentence to its corresponding evidence sentence indices.

### 3.5. Implementation Details

All experiments were conducted in Python via the Gemini 2.5 Pro API. For tasks requiring strict JSON compliance (Subtasks 1, 2, and 4), the API temperature was set to absolute zero to prevent variation. We slightly increased this to 0.1 for Subtask 3 to help the generated answers read more naturally. Additionally, to ensure the model wouldn't censor valid clinical terminology, we bypassed the standard API safety blocks by setting every `HarmCategory` to `BLOCK_NONE`.

### 3.6. Evaluation Metrics

We evaluated our system using the official metrics defined by the ArchEHR-QA organizers. For the generative tasks (Subtasks 1 and 3), performance was measured using a mix of lexical, semantic, and clinical scores. Specifically, both subtasks relied on ROUGE<sub>L</sub>sum (Lin, 2004), BERTScore (Zhang et al., 2019), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023) to ensure clinical concepts were preserved accurately. Subtask 3 also included BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016) to evaluate how well the model simplified the text. The official overall score for both generative subtasks is calculated as the unweighted average of their respective constituent metrics.

For the information extraction and classification components (Subtasks 2 and 4), the evaluation relies on standard precision, recall, and F1-scores. Subtask 2 is evaluated comprehensively under both strict and lenient boundary-matching criteria, reporting these metrics at both the micro and macro levels; the strict micro F1-score is designated by the organizers as the overall ranking metric for this subtask. Similarly, Subtask 4 is assessed utilizing micro precision, micro recall, and micro F1-scores. In strict alignment with the official task guidelines, the micro F1-score serves as the overall aggregate metric for Subtask 4.

## 4. Result Analysis

In this section, we analyze the performance of the HealthNLP\_Retrievers pipeline on the ArchEHR-QA 2026 dataset. Our evaluation follows the dual-metric framework of the challenge, focusing on factuality (alignment and evidence identification) and relevance (textual quality) (Soni and Demner-Fushman, 2026b). As shown in Table 1, our cascaded system demonstrated competitive performance across all four tracks, validating the efficacy of our multi-stage LLM prompting strategy.

| Subtask                 | Rank | Primary Metric  | Ours | Top  | Secondary Metric | Ours | Top  |
|-------------------------|------|-----------------|------|------|------------------|------|------|
| Question Interpretation | 1    | Overall         | 31.2 | 31.2 | BERTScore        | 46.8 | 46.8 |
| Evidence Scoring        | 7    | Strict Micro F1 | 60.2 | 63.7 | Lenient Micro F1 | 68.3 | 72.4 |
| Answer Generation       | 5    | SARI            | 59.2 | 58.6 | AlignScore       | 33.6 | 31.7 |
| Alignment               | 9    | Micro F1        | 76.9 | 81.5 | Micro Precision  | 83.8 | 88.0 |

Table 1: Official leaderboard performance of the HealthNLP\_Retrievers system compared to the top-performing system in each subtask.

#### 4.1. Interpretative Accuracy

Our few-shot query interpretation module achieved the 1st place ranking on the ArchEHR-QA leaderboard for Subtask 1, earning an overall score of 31.2. The system demonstrated strong performance across both lexical and semantic evaluations, recording a ROUGE<sub>L</sub>sum score of 35.3 and a BERTScore of 46.8. This secured the top position in a highly competitive field, representing a narrow but decisive improvement over the runner-up team, KPSCMI (overall score of 30.8), and the third-place team, OptiMed (overall score of 29.9).

This outcome highlights the real-world effectiveness of our persona-based prompting strategy. By instructing the model to act as a clinical administrative assistant and enforcing a strict 15-word limit, we successfully stripped away the emotional, verbose parts of the patient’s question to leave a tight clinical search query. While other teams likely struggled to cut the word count without dropping critical details, our BERTScore of 46.8—which sits noticeably higher than KPSCMI’s 41.0 and OptiMed’s 43.1—proves we could aggressively summarize requests without losing their actual meaning. Keeping that core intent intact gave our downstream retrieval module a solid, objective clinical target to search for, rather than getting thrown off by informal patient phrasing.

#### 4.2. Evidence Identification

In the evidence extraction phase, our 1–5 Likert-scale scoring approach achieved a strict micro F1 score of 60.2, placing us 7th overall (compared to the top system, Neural, at 63.7). Looking closely at the metrics, our strict micro recall (62.1) noticeably outperformed our strict micro precision (58.4). This result was entirely intentional. We specifically designed the retrieval prompt to lean heavily toward recall to ensure the downstream generator never suffered from evidence starvation. We knew this safety-first approach would cost us some strict precision compared to the top teams, as the model occasionally flagged helpful background context as essential evidence.

The lenient evaluation scores clearly back up this design trade-off. Under the lenient framework, our micro precision jumped 17.3 points to 75.7, pulling

our overall lenient micro F1 up to 68.3. This massive gap between the strict and lenient precision tells us exactly what the model was doing: while it pulled in too much text according to the rigid gold-standard boundaries, those extra sentences were still highly relevant and clinically adjacent to the actual answer. By accepting a few more false positives, we successfully kept our false negative rate low. This gave the final text generation stage a much richer set of facts to work with, rather than filling the context window with completely irrelevant noise.

#### 4.3. Grounded Generation

The generation module achieved 5th place on the ArchEHR-QA leaderboard with an overall score of 34.6. The system’s performance was anchored by a strong SARI score of 59.2 and a BERTScore of 43.8. Notably, our SARI score—a metric specifically designed to evaluate text rewriting and simplification—outperformed several higher-ranked teams, including the 1st-place team, WisPerMed (58.6), and the 4th-place team, Neural (57.7). This underscores the effectiveness of our clinical documentation specialist persona and soft-cut truncation heuristic in synthesizing concise, professionally formatted responses that effectively translate complex clinical context into accessible language.

Despite these strong semantic indicators, the relatively low BLEU score of 7.0 highlights a universal vocabulary gap challenge in clinical text generation, mirroring the single-digit BLEU scores observed even among the top-performing teams (e.g., 9.9 for WisPerMed). While our system generated semantically accurate professional responses, the specific lexical phrasing often diverged from the rigid n-gram overlaps expected by the reference gold standard. Crucially, the MEDCON score of 38.7 indicates a robust adherence to medical concept constraints. This proves that our strict RAG-based architecture successfully restricted its generation strictly to the provided evidence pool, synthesizing the necessary clinical information without hallucinating external or unsupported medical concepts.

| Subtask 1       |           |           |            |               |
|-----------------|-----------|-----------|------------|---------------|
| Setting         | BERTScore | ROUGELsum | AlignScore | MEDCON (UMLS) |
| Zero-shot       | 34.13     | 18.75     | 12.80      | 9.20          |
| Few-shot (Ours) | 45.09     | 32.55     | 23.10      | 17.20         |

| Subtask 2              |                 |                        |                     |                  |
|------------------------|-----------------|------------------------|---------------------|------------------|
| Anchor                 | Strict Micro F1 | Strict Micro Precision | Strict Micro Recall | Lenient Micro F1 |
| Patient Narrative      | 60.20           | 50.56                  | 74.38               | 63.60            |
| Clinician Query (Ours) | 62.90           | 61.42                  | 64.46               | 65.82            |

Table 2: Impact of few-shot formulation (Subtask 1) and clinical query anchors (Subtask 2).

#### 4.4. Evidence Alignment

Our alignment module finished in 9th place for Subtask 4 with an overall micro F1 of 76.9. We specifically designed the prompt instructions to be highly selective, which pushed the model to only extract the most certain evidence. This precision-heavy approach resulted in a micro precision score of 83.8. We found this result particularly competitive, as it slightly outperformed or matched top-tier systems like Yale-DM-Lab (83.3) and UIC-AIHealth4All (83.6). These numbers confirm that the module is highly effective at filtering out irrelevant noise and isolating only the most direct supporting sentences from the notes.

On the other hand, our micro recall of 71.1 reflects a specific design choice in our architecture. Because we used such strict prompting heuristics, the model occasionally skipped over sentences that provided implicit support or were part of complex, multi-hop evidence chains. However, we view this high-precision focus as a necessary safeguard for a rigorous verification layer. By favoring exact evidence over a wider contextual window, we ensured the system only cited facts that were explicitly stated in the clinical note. This essentially acts as a final check to prevent the model from making up connections that aren't explicitly stated in the clinical record.

#### 4.5. Ablation Study

To evaluate the impact of our primary design choices, we conducted ablation experiments across Subtask 1 and Subtask 2 using the official development dataset, with the results summarized in Table 2. For the question interpretation task (Subtask 1), we tested our few-shot clinical assistant persona against a standard zero-shot baseline. The few-shot setup delivered major improvements across all semantic and clinical metrics. Most notably, the MEDCON score jumped from 9.20 to 17.20. This significant increase in medical entity retention shows that using curated examples is essential for bridging the gap between informal patient

language and professional clinical terms.

For evidence identification (Subtask 2), we looked at whether using reformulated clinician queries worked better than using the raw patient narratives as retrieval anchors. While the longer patient narratives naturally led to a higher recall (74.38), they also introduced a lot of noise, which dragged precision down to 50.56. In contrast, our clinician-centered queries achieved a much better balance, raising the Strict Micro F1 to 62.90 and Precision to 61.42. This confirms that a concise, 15-word query is much more effective at filtering out irrelevant patient details while keeping the core clinical facts needed for accurate evidence retrieval.

#### 4.6. Discussion

Our results highlight a clear tension between making a query concise and keeping it clinically detailed. In Subtask 1, while we took first place overall, our MEDCON score of 18.7 lagged behind KPSCMI (27.9) and Neural (25.6). This suggests that our administrative assistant persona and 15-word limit were almost too effective at cutting noise—they sometimes stripped away critical medical entities in favor of general descriptions. For example, when a patient described a “dull to deep pain” while listing a toxicological history of “trihexyphenidyl, thorazine, and cocaine,” the gold-standard query filtered the descriptive fluff but kept the specific drugs. Our system, however, generalized this into: “What are the potential causes of the patient’s persistent chest pain?” While our version was readable, it lost the specific toxicological context that a human clinician would immediately flag as a high-priority lead. This preference for accessible, general language over technical precision also impacted Subtask 3; while our high SARI scores confirm the model successfully improved text simplicity for a lay audience, this same behavior resulted in a low BLEU score (7.0). This indicates that while our generative outputs remained semantically accurate, the stylistic shift toward simplified vocabulary created a disconnect with the exact n-gram overlaps of the more

technically-worded reference labels.

Beyond these vocabulary issues, we saw a clear imbalance between precision and recall in Subtasks 2 and 4. In Subtask 2, our recall-heavy approach pushed precision down to 58.4, while the precision-focused prompting in Subtask 4 limited our recall to 71.1, falling behind teams like OptiMed (79.8). It appears our alignment layer became too exclusionary, often discarding implicit support to verify core facts. In Case 4, for instance, our model correctly identified that the patient had “low-output, acute-on-chronic heart failure.” While the system verified the specific procedures and outcomes (Sentences 5, 6, 13, 19, 20), it missed the foundational admission context and diagnostic proof, such as the 25% ejection fraction and congestive hepatopathy (Sentences 10, 11, 18). This shows that while strict filtering is great for verifying specific treatments, it can severely hurt recall by stripping away the underlying diagnostic reasoning that clinicians rely on.

## 5. Conclusion

We presented the HealthNLP\_Retrievers system, a cascaded pipeline for the ArchEHR-QA 2026 shared task. By decomposing clinical QA into query interpretation, heuristic scoring, and precise alignment, our system effectively bridged the gap between patient narratives and clinical notes. This modular design, driven by our few-shot query reformulation strategy, secured 1st place in Question Interpretation (Subtask 1) with an overall score of 31.2 and established a competitive framework for grounded evidence extraction.

Our ablation analysis demonstrated that formulating concise clinical queries significantly improves medical entity retention and retrieval precision. However, error analysis revealed a continuing performance trade-off between concise summarization and medical entity retention, as evidenced by our MEDCON scores. Furthermore, our high-precision alignment strategy in Subtask 4 successfully mitigated hallucinations but resulted in lower recall compared to top-tier competitors. Future work will focus on iterative self-correction to reduce error propagation and the distillation of this pipeline into open-weight models for privacy-preserving clinical deployment.

## 6. Limitations

This work has several limitations. First, the system relies exclusively on the Gemini 2.5 Pro API, introducing a dependency on a proprietary, closed-weight model. This limits reproducibility and adaptability in privacy-sensitive environments, as all clinical text is transmitted to an external server. Sec-

ond, our evaluation is confined to the ArchEHR-QA 2026 benchmark; performance on other clinical QA datasets or EHR systems remains unexplored. Third, the strictly cascaded architecture propagates errors across stages: an imprecise query reformulation in Subtask 1 can degrade evidence retrieval precision in Subtask 2, which in turn may degrade the quality of the generated answer. Future work should address these gaps through evaluation on open-weight, locally deployable models and the development of end-to-end strategies to mitigate cascaded error propagation.

## 7. Ethics Statement

This work utilizes the ArchEHR-QA dataset (Soni and Demner-Fushman, 2026a), which is derived from the de-identified MIMIC-III database (Johnson et al., 2016) and accessed through PhysioNet under appropriate data use agreements. To mitigate privacy risks, the patient questions in the dataset are inspired by, rather than directly sourced from, real clinical inquiries. Furthermore, in compliance with both the MIMIC-III data use agreement and ArchEHR-QA 2026 shared task guidelines, we ensured that only strictly de-identified clinical note excerpts were transmitted to the Gemini API. No raw patient identifiers were included in any API calls, and all data processing remained fully compliant with PhysioNet regulations.

We emphasize that the system described herein is designed exclusively as a research prototype and is not intended for direct clinical deployment without rigorous validation by qualified healthcare professionals. For the purposes of this experimental evaluation, all model safety filters were disabled solely to allow the system to accurately process explicit clinical terminology without unwarranted refusal.

## 8. Bibliographical References

- Asma Ben Abacha and Dina Demner-Fushman. 2019a. Bridging the gap between consumers’ medication questions and trusted answers. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 25–29. IOS Press.
- Asma Ben Abacha and Dina Demner-Fushman. 2019b. On the role of question summarization and information source restriction in consumer health question answering. *AMIA Summits on Translational Science Proceedings*, 2019:117.
- Tal Burt, KS Button, HHZ Thom, RJ Noveck, and Marcus R Munafò. 2017. The burden of the “false-negatives” in clinical development: Analyses of



- current and alternative scenarios and corrective measures. *Clinical and translational science*, 10(6):470–479.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- 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.
- Zainab Ghafoor, Md Shafiqul Islam, Koushik Howlader, Md Rasel Khondokar, Tanusree Bhattacharjee, Sayantan Chakroborty, Adrito Roy, Ushashi Bhattacharjee, and Tirtho Roy. 2025. Enhancing medical ai safety through multi-agent evaluation and iterative refinement. In *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–1.
- 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.
- 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):160035.
- Vojtech Lanz and Pavel Pecina. 2025. CUNI-a at ArchEHR-QA 2025: Do we need giant LLMs for clinical QA? In *Proceedings of the 24th Workshop on Biomedical Language Processing*.
- 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.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Liqiang Nie, Yi-Liang Zhao, Mohammad Akbari, Jialie Shen, and Tat-Seng Chua. 2014. Bridging the vocabulary gap between health seekers and healthcare knowledge. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):396–409.
- 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*.
- 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.
- 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.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950.
- Sarvesh Soni and Dina Demner-Fushman. 2026a. A dataset for addressing patient’s information needs related to clinical course of hospitalization. *Scientific Data*.
- 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.
- 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.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer,

- Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Wenxuan Wang, Zizhan Ma, Meidan Ding, Shiyi Zheng, Shengyuan Liu, Jie Liu, Jiaming Ji, Wenting Chen, Xiang Li, Linlin Shen, et al. 2025. Medical reasoning in the era of llms: a systematic review of enhancement techniques and applications. *arXiv preprint arXiv:2508.00669*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed 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, pages 24824–24837.
- Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific.
- 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.
- 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.
- Seksan Yoadsanit, Nopporn Lekuthai, Watcharitpol Sermsrisuwan, and Titipat Achakulvisut. 2025. LAMAR at ArchEHR-QA 2025: Clinically aligned LLM-generated few-shot learning for EHR-grounded patient question answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing*.
- 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.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

# Yale-DM-Lab at ArchEHR-QA 2026: Deterministic Grounding and Multi-Pass Evidence Alignment for EHR Question Answering

Elyas Irankhah, Samah Fodeh

Yale University, Department of Emergency Medicine  
New Haven, CT, USA

Elyas.irankhah@yale.edu, samah.fodeh@yale.edu

## Abstract

We describe the Yale-DM-Lab system for the ArchEHR-QA 2026 shared task. The task studies patient-authored questions about hospitalization records and contains four subtasks (ST): clinician-interpreted question reformulation, evidence sentence identification, answer generation, and evidence-answer alignment. ST1 uses a dual-model pipeline with Claude Sonnet 4 and GPT-4o to reformulate patient questions into clinician-interpreted questions. ST2–ST4 rely on Azure-hosted model ensembles (o3, GPT-5.2, GPT-5.1, and DeepSeek-R1) combined with few-shot prompting and voting strategies. Our experiments show three main findings. First, model diversity and ensemble voting consistently improve performance compared to single-model baselines. Second, the full clinician answer paragraph is provided as additional prompt context for evidence alignment. Third, results on the development set show that alignment accuracy is mainly limited by reasoning. The best scores on the development set reach 88.81 micro F1 on ST4, 65.72 macro F1 on ST2, 34.01 on ST3, and 33.05 on ST1.

**Keywords:** EHR question answering, Evidence alignment, LLM ensemble, Few-shot prompting

## 1. Introduction

Electronic health records (EHRs) contain dense clinical narratives. They are difficult for patients to interpret without clinical training. Prior work has studied information extraction from EHR text (Liu et al., 2013) and clinical question answering with large language models (LLMs) (Singhal et al., 2023; Nori et al., 2023). The ArchEHR-QA shared task (Soni and Demner-Fushman, 2026b) targets patient-facing access. It asks systems to answer patient-authored questions about a hospitalization episode. The task has four linked subtasks.

A system must reformulate a patient question into a clinician-interpreted question, identify supporting note sentences, generate a grounded answer, and align each answer sentence to its evidence. Evidence grounding is central. Retrieval-augmented generation is a common mechanism for conditioning generation on external text (Lewis et al., 2020). Citation attribution and faithfulness evaluation provide complementary ways to assess whether model outputs are supported by sources (Gao et al., 2023; Honovich et al., 2022).

This paper describes the Yale-DM-Lab submission. Our approach is prompt-based and utilizes in-context learning (Brown et al., 2020). We use self-consistency and voting to improve robustness (Wang et al., 2023). We also evaluate reasoning-style prompting variants when appropriate (Wei et al., 2022). ST1 uses Claude Sonnet 4 and GPT-4o via their respective direct APIs. ST2–ST4 use Azure-hosted ensembles (o3, GPT-5.2, GPT-5.1, and DeepSeek-R1). We focus on controlled abla-

tions over ensemble size, few-shot count, and vote thresholds.

Our contributions are:

- A modular four-subtask pipeline for EHR question answering with a shared Azure infrastructure and consistent few-shot prompting design.
- A systematic ablation study evaluating ensemble size, few-shot count, and voting thresholds for evidence identification and alignment.
- An analysis of evidence grounding in ST4 showing that providing the full clinician answer paragraph as context improves evidence alignment.

## 2. Related Work

Information extraction from electronic health record (EHR) text has long been studied in clinical natural language processing (NLP) (Liu et al., 2025). Early work by Demner-Fushman et al. (2009) developed clinical decision support methods based on structured information extracted from clinical narratives. More recent datasets extend this direction toward question answering. Kweon et al. (2024) introduced patient-specific QA benchmarks grounded in discharge summaries which enabled evaluation of models on real clinical documentation. Large language models (LLMs) have recently advanced medical question answering. Roberts et al. (2020) demonstrated that large-scale pretrained models acquire substantial factual medical knowledge through language modeling. Building on this

progress, [Lehman et al. \(2023\)](#) evaluated LLMs in clinical reasoning settings and showed strong performance across medical QA tasks. However, these studies also highlighted limitations when models operate on long clinical documents. EHR-based QA requires reasoning over dense terminology and extended context. Answers must remain grounded in specific source sentences rather than general knowledge. Several studies also address grounding through retrieval mechanisms. [Izacard and Grave \(2021\)](#) proposed retrieval-augmented generation (RAG), which conditions generation on retrieved passages and improves factual consistency. In clinical domains, grounding is especially important because hallucinated statements may introduce clinical risk. One study by [Ji et al. \(2023\)](#) analyzed hallucination behaviors in LLMs and showed that unsupported generations remain a persistent problem. Complementary work by [Rashkin et al. \(2021\)](#) examined attribution and truthfulness that emphasized the importance of linking generated claims to supporting evidence. Prompting strategies have been explored to improve reasoning reliability without additional training. [Saxena et al. \(2024\)](#) showed that few-shot prompting with labeled examples improves reasoning consistency even without fine-tuning. [Zhou et al. \(2023\)](#) presented chain-of-thought prompting to encourage intermediate reasoning steps. Extending this idea, [Fu et al. \(2023\)](#) proposed self-consistency decoding, which samples multiple reasoning paths and selects the most stable answer. More recently, [Kuligin et al. \(2025\)](#) demonstrated that ensembles of diverse models reduce correlated reasoning errors by combining complementary inductive biases. Evaluation of generated clinical text has also evolved. [Kryściński et al. \(2019\)](#) showed that lexical overlap metrics alone fail to capture factual correctness in abstractive summarization. [Xu et al. \(2016\)](#) introduced SARI to better evaluate paraphrased outputs. Semantic metrics such as BERTScore further improve evaluation of meaning preservation [Zhang et al. \(2020\)](#). These metrics are particularly relevant for patient-facing clinical answers, which typically paraphrase source text. BLEU therefore provides only limited signal in clinical lay-language generation.

### 3. Shared Task and Dataset

The ArchEHR-QA 2026 dataset contains 167 cases derived from de-identified EHR narratives ([Soni and Demner-Fushman, 2026a](#)). Each case includes a patient free-text question, a clinical note segmented into numbered sentences, a clinician-interpreted question, a clinician answer paragraph, and gold evidence alignments. The dataset is split into 20 development cases and 147 test cases. Subtask

1 reformulates the patient question into a clinician-interpreted question of at most 15 words and is evaluated using ROUGE-Lsum, BERTScore, AlignScore, and MEDCON, with the leaderboard score computed as their mean. Subtask 2 identifies the minimal set of supporting sentence IDs and is evaluated using strict micro F1. Subtask 3 generates an answer of at most 75 words grounded in the provided clinical note and is evaluated using BLEU, ROUGE, SARI, BERTScore, AlignScore, and MEDCON. Subtask 4 aligns each clinician answer sentence with supporting note sentences and is evaluated using micro F1. MEDCON is part of the official evaluation for Subtasks 1 and 3 but could not be reproduced locally due to restricted access to the official evaluation environment. Therefore, development experiments reported in this paper use locally reproducible metrics only.

## 4. Methodology

### 4.1. Prompting Setup

We use task-specific prompt templates with fixed instruction blocks. We vary (i) few-shot example count, (ii) model ensemble composition, and (iii) vote threshold and post-processing settings. Few-shot examples are selected in a leave-one-out manner from development cases and are excluded if they match the current case. Unless stated otherwise, prompts enforce strict JSON outputs for structured tasks (ST2, ST4) and task constraints such as word limits and no unsupported facts for generative tasks (ST1, ST3). Across subtasks, the core instruction text remains fixed while experimental variations are applied to configuration parameters. For ST2, we vary few-shot count (3, 10, 19), ensemble members, and merge thresholds. For ST3, we fix the two-stage faithful scaffold and vary few-shot count and model configuration. For ST4, we keep the alignment instruction fixed and vary ensemble and self-consistency settings, vote thresholds, and recall augmentation.

### 4.2. Subtask 1 – Question Reformulation

ST 1 uses Claude Sonnet 4 and GPT-4o via their direct APIs to reformulate the patient question into a  $\leq 15$ -word clinician-interpreted question. We first evaluate a single GPT-4o baseline. We then introduce a dual-model pipeline where Claude Sonnet 4 (Anthropic API) and GPT-4o (OpenAI API) run in parallel. We use:

- **Configurations:** single-model GPT-4o baseline, base dual-model run, and a clinical-reasoning variant with an explicit chain-of-thought step.

- **Dual-model setup:** Claude Sonnet 4 and GPT-4o in parallel; merge by rule selector (pattern match on clinical intent) with fallback to pattern-alignment score.
- **Prompt pipeline (three stages):** (1) analyze the clinical note for procedures, medications, findings, and diagnoses to ground the reformulation; (2) retrieve up to five dev-set examples via hybrid score (question-type + lexical similarity) as few-shot; (3) score generated candidates with gold question-style patterns and select the highest. Gold questions refer to the ground-truth clinician-interpreted questions from the development set. Candidate questions are scored using pattern alignment heuristics (e.g., question-type matching and lexical similarity to dev gold templates), and the highest-scoring candidate is selected.
- **Hard constraints:** at most 15 words, no first-person pronouns. Figure 1 shows the overall architecture of the Subtask 1 dual-LLM reformulation pipeline.

Clinical note excerpts are used only during development. For the official Subtask 1 test set, note excerpts are not provided. The pipeline therefore operates in a patient-question-only setting, and note-dependent components are not used. All reported test results for Subtask 1 are obtained without clinical notes. Clinical context extraction is a prompt-based preprocessing step that identifies key clinical elements explicitly stated in the inputs. The inputs include the patient question and, during development, an optional note excerpt. Extracted elements include procedures, medications, diagnoses, findings, and temporal or urgency cues. This step is implemented using an LLM instruction prompt, not a separate trained model. The extracted elements guide candidate generation and constrain the reformulation toward the intended clinical information need. Only explicitly stated information is used, and no new clinical facts are introduced.

#### 4.3. Subtask 2 – Evidence Identification

The system prompt instructs the model to return a JSON array of sentence IDs that constitute the minimal evidence set. We explored:

- **Vote threshold:** a sentence is included if it receives  $\geq k$  votes across  $n$  model-run pairs. Setting  $k = 1$  (union) maximizes recall;  $k = \lceil n/2 \rceil + 1$  (majority) maximizes precision.
- **Few-shot count:** we use 3, 10, or 19 leave-one-out development examples. For each test case, examples are selected from other development cases (excluding the current case)

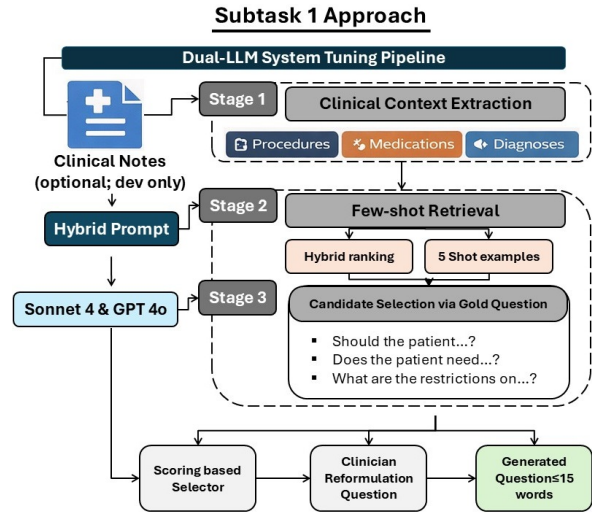


Figure 1: Dual-LLM pipeline for Subtask 1 clinician-interpreted question reformulation. The system extracts clinical context, retrieves few-shot examples, generates candidates using parallel LLMs, and selects the final question via scoring-based constraint enforcement. Clinical notes are optional and used only during development; test inference uses patient questions only.

and include the patient question, clinician-interpreted question, and gold evidence sentence IDs. Only cases with non-empty gold evidence are used.

- **Ensemble models:** o3 + GPT-5.2 + GPT-5.1 (baseline trio), with optional addition of DeepSeek-R1.
- **Contrast few-shot:** GOOD examples (correct gold evidence sets) plus BAD examples, e.g. one over-inclusive example where the gold set is augmented with a spurious sentence ID so the model sees what to avoid. For instance, a GOOD set may be ["2", "5"], while a BAD set may be ["2", "5", "9"], where sentence "9" is irrelevant. Preference-based methods such as DPO (Rafailov et al., 2024; Fodeh et al., 2026) use chosen vs. rejected pairs for *training*; we apply the same idea only in-context.

The best configuration uses 10 few-shot examples with union merge ( $\text{min\_votes} = 1$ ) and achieves 64.17 strict micro F1 on the development set. Post-processing drops invalid sentence IDs (not present in the case), removes duplicates, and sorts IDs. In the enhanced configuration, we apply additional conservative post-processing (e.g., filtering low-confidence candidates) to improve the score to 65.33.



#### 4.4. Subtask 3 – Answer Generation

The prompt gives the patient question, the clinician-interpreted question, the identified evidence sentences (from ST2), and the full note. The model produces a free-text answer, which we hard-truncate to 75 words. We compare:

- **Single model baseline:** o3 only (zero temperature, 5-shot).
- **Ensemble:** o3 + GPT-5.2 + GPT-5.1 with 10 or 15 few-shot dev examples.
- **Faithful variants:** two-stage pipelines that first generate a cited evidence-grounded draft and then rewrite the answer using only the cited sentences. Stage 2 is constrained to avoid unsupported content and not introduce new facts. We evaluate both single-model (GPT-5.2) and ensemble configurations. For the ‘Faithful ensemble + BERT selection’ setting, multiple candidate answers are generated from the ensemble, and the final answer is selected by BERTScore-based reranking against the case note text (used as a proxy target).
- **Few-shot ablations:** we evaluate 5, 10, and 15 dev examples to study prompt sensitivity.

*In the faithful two-stage setting, Stage 1 produces a draft answer with inline citations (e.g., [12], [15]) that point to the provided evidence sentences. Stage 2 rewrites the answer using only the cited evidence, removes citation markers, and enforces the 75-word limit. This constrains the output to preserve clinical entities and avoids unsupported additions. Example (schematic): Input: question + evidence ([12], [15]). Stage 1: "... [12][15]". Stage 2: "...". The final output preserves the same facts and remains supported by the cited evidence.*

#### 4.5. Subtask 4 – Evidence Alignment

ST4 is the most extensively studied subtask as shown in Figure 2. The prompt includes the patient question, the clinician-interpreted question, the full note (numbered sentences), and all answer sentences (numbered). We use:

- **Ensemble and self-consistency:**  $M$  deployments, each with  $S$  self-consistency samples at  $T_s \in \{0.3, 0.4\}$ , plus one run at  $T = 0$ . For o3 we use  $T = 1.0$ . A link is kept if its vote count  $\geq \theta$ , with  $\theta = k$  (manual override) or  $\theta = \lfloor MS/2 \rfloor + 1$  (majority). Here,  $MS$  is the total number of votes ( $M$  models  $\times S$  self-consistency samples each), we aggregate votes at the (answer\_id, evidence\_id) link level, keep links supported by more than half of all votes under majority voting, and use  $k$  as a manual precision/recall trade-off.

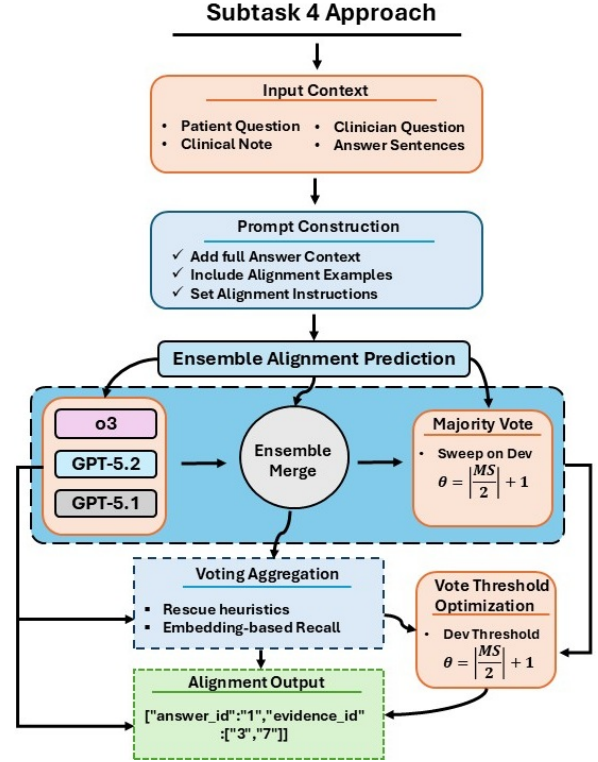


Figure 2: Pipeline for Subtask 4 evidence alignment. The system combines ensemble inference, voting aggregation, and embedding-based recall to improve evidence matching.

- **Dev threshold sweep:** On dev we sweep  $\theta \in \{1, \dots, MS\}$ , pick the value that maximizes micro F1, write it to `best_vote_threshold.txt`, and reuse it at test time.
- **Full-answer context:** The `clinician_answer_without_citation` paragraph from the key is added as a “Full clinician answer (for context)” block so the model can resolve anaphora and follow the narrative; earlier we used only isolated answer sentences.
- **Embedding-based recall augmentation:** Additional (answer, note) pairs are introduced when their sentence-transformer similarity exceeds a threshold  $\tau$  (default 0.68). This mechanism acts as a post-vote recall step to recover missed links after ensemble aggregation. In this work, this step corresponds to the “rescue heuristics” reported in the results.

## 5. Results

Table 1 reports performance across all four subtasks. We show development experiments and final test submissions. The experiments evaluate

ensemble size, few-shot count, vote thresholds, and grounding strategies.

**Subtask 1: Question Reformulation.** The best development configuration uses note grounding and pattern-based scoring. It achieves a score of 33.05. Dual-model generation improves performance over a single GPT-4o baseline. Chain-of-thought prompting does not provide consistent gains. On the test set, the same configuration achieves 27.09. The decrease reflects the difficulty of producing short clinician-interpreted questions under strict lexical constraints.

**Subtask 2: Evidence Identification.** Evidence identification benefits from ensemble voting. Majority voting increases precision but lowers recall. Union merging increases recall. This improves micro F1. The best development configuration uses 10 few-shot examples with union merging. It achieves 64.17 micro F1. Post-processing increases the score to 65.33. On the test set, the same configuration reaches 61.90 micro F1. These results show that recall-oriented aggregation improves evidence retrieval.

**Subtask 3: Answer Generation.** Answer generation remains difficult. BLEU scores are low. Clinical answers often use varied wording. Ensemble generation improves the overall score compared to single-model generation. The best development score is 34.01. This result uses the o3, GPT-5.2, and GPT-5.1 ensemble. The faithfulness pipeline achieves similar performance. On the test set, the ensemble reaches 30.95.

**Subtask 4: Evidence Alignment.** Evidence alignment produces the strongest results. Ensemble voting improves stability. Embedding-based recall augmentation (rescue heuristics) increases recall. The best development configuration reaches 88.81 micro F1. This setup uses ensemble aggregation with recall-oriented augmentation. Self-consistency sampling provides limited improvement, with performance reaching 83.39 micro F1, which remains below recall-augmented configurations. For example, Table 1 shows 83.39  $\mu$ F1 for the self-consistency configuration versus 88.81  $\mu$ F1 with embedding-based recall augmentation. On the test set, the ensemble with rescue heuristics achieves 80.41 micro F1.

**Observations.** Several consistent patterns emerge across the experiments. Model diversity provides the most reliable improvement across subtasks. Ensembles outperform single-model configurations and produce more stable predictions. Ensemble disagreement is handled through

sentence/link-level vote aggregation. Disagreements occur mainly on borderline evidence, where union-based aggregation preserves recall while majority voting improves precision but may omit relevant sentences. This effect is most visible in ST2 and ST4, where ensemble voting improves robustness and reduces prediction variance. Few-shot prompting also improves performance when examples are well selected. Increasing the number of examples helps until results stabilize. This suggests that example quality is more important than quantity. Aggregation strategy strongly affects evidence retrieval. Recall-oriented approaches perform better than precision-oriented voting. In ST2, union-based merging consistently achieves higher micro F1 than strict majority voting. Missing relevant evidence appears more harmful than including extra candidate sentences, as observed in ST2, where recall-oriented union merging achieves higher micro F1 than precision-oriented majority voting (64.17 vs. 53.40). This effect is not directly measured in ST3 but likely propagates to answer generation through upstream evidence selection. Contextual grounding improves evidence alignment. Providing the full clinician answer paragraph reduces ambiguity and improves matching between answer sentences and supporting note sentences. Evidence alignment remains the most stable task in the pipeline, while answer generation remains the most challenging. Multi-step reasoning continues to introduce errors even for large language models. In ST3, errors arise when combining multiple evidence sentences leads to unsupported or loosely grounded paraphrases. In ST4, errors appear as missing or incorrect alignment links when reasoning fails to consistently map answer statements to all supporting evidence.

## 6. Conclusion

We presented the Yale-DM-Lab system for ArchEHR-QA 2026, a four-subtask pipeline for patient-facing EHR question answering. The system combines model-diverse LLM ensembles, self-consistency voting, and leave-one-out few-shot prompting. Our experiments present strong performance across subtasks, with evidence alignment (ST4) reaching 88.81 micro F1 on the development set. The full clinician answer paragraph is used as additional context during evidence alignment. The result emphasizes the critical role of contextual grounding in evidence-based clinical question answering. The implementation of our system, including prompts and inference scripts for all subtasks, is publicly available at <https://github.com/Data-Mining-Lab-Yale/ArchEHR2026-CL4Health-LREC>.

| Development Phase                              |              |         |              |       |       |              |
|------------------------------------------------|--------------|---------|--------------|-------|-------|--------------|
| Method / Config                                | Metrics      |         |              |       |       |              |
| <i>Subtask 1 — Question Reformulation</i>      |              |         |              |       |       |              |
|                                                | Score        | R1      | R2           | RLsum | BERT  | Align        |
| Single (GPT-4o) + 5-shot <sup>‡</sup>          | 25.49        | 34.67   | 15.28        | 31.76 | 44.71 | —            |
| Dual-model ensemble <sup>‡</sup>               | 26.14        | 36.28   | 16.52        | 33.06 | 45.34 | —            |
| Dual + CoT + note grounding <sup>†</sup>       | 23.98        | 32.33   | 12.48        | 29.19 | 42.74 | —            |
| <b>Dual + note grounding + pattern scoring</b> | <b>33.05</b> | 46.35   | 29.28        | 44.57 | 54.58 | —            |
| Dual-LLM reasoning                             | 32.14        | 36.28   | 16.52        | 33.06 | 45.34 | 23.85        |
| <i>Subtask 2 — Evidence Identification</i>     |              |         |              |       |       |              |
|                                                | $\mu$ P      | $\mu$ R | $\mu$ F1     | mP    | mR    | mF1          |
| 10-shot, majority                              | 72.86        | 42.15   | 53.40        | 55.20 | 41.76 | 45.32        |
| 19-shot, majority                              | 70.83        | 42.15   | 52.85        | 53.32 | 41.66 | 44.53        |
| 3-shot, union                                  | 63.55        | 56.20   | 59.65        | 62.40 | 53.84 | 54.07        |
| 10-shot, union                                 | 64.71        | 63.64   | 64.17        | 58.85 | 59.55 | 56.77        |
| 19-shot, union                                 | 61.54        | 59.50   | 60.50        | 57.12 | 57.33 | 55.33        |
| <b>10-shot, union + post-proc</b>              | 54.75        | 80.99   | 65.33        | 58.74 | 81.88 | <b>65.72</b> |
| 3-shot + contrast few-shot, union              | 57.89        | 63.64   | 60.63        | 52.28 | 59.13 | 52.78        |
| 4-model (+R1), 10-shot, union                  | 54.66        | 72.73   | 62.41        | 54.29 | 72.12 | 60.03        |
| <i>Subtask 3 — Answer Generation</i>           |              |         |              |       |       |              |
|                                                | Score        | BLEU    | RLsum        | SARI  | BERT  | Align        |
| Single (o3), 5-shot                            | 27.42        | 2.36    | 17.05        | 55.75 | 34.51 | —            |
| Two-stage evidence → answer (GPT-5.2)          | 33.62        | 9.33    | 26.21        | 58.54 | 40.34 | 27.60        |
| Faithful ensemble + BERT selection             | 33.99        | 8.95    | 26.74        | 58.01 | 41.71 | 27.33        |
| <b>o3 + GPT-5.2 + GPT-5.1</b>                  | <b>34.01</b> | 9.78    | 26.28        | 58.16 | 39.69 | 28.82        |
| Ensemble + faithfulness, 10-shot               | 33.53        | 9.81    | 25.16        | 59.18 | 39.98 | —            |
| Ensemble, 10-shot                              | 28.28        | 2.40    | 18.59        | 56.49 | 35.64 | —            |
| Ensemble, 15-shot                              | 27.11        | 2.63    | 17.25        | 53.79 | 34.77 | —            |
| <i>Subtask 4 — Evidence Alignment</i>          |              |         |              |       |       |              |
|                                                | $\mu$ P      | $\mu$ R | $\mu$ F1     | mP    | mR    | mF1          |
| GPT-5.2 + GPT-5.1, 10-shot                     | 88.41        | 88.41   | 88.41        | 90.54 | 88.73 | 88.50        |
| <b>Embedding recall augmentation</b>           | 88.49        | 89.13   | <b>88.81</b> | 90.13 | 90.04 | 89.43        |
| Ensemble + Self-consistency (SC)               | 84.96        | 81.88   | 83.39        | 87.89 | 84.96 | 85.12        |
| Embedding-only                                 | 65.91        | 42.03   | 51.33        | 74.21 | 44.40 | 52.16        |
| Test Phase                                     |              |         |              |       |       |              |
| Submission                                     | Metrics      |         |              |       |       |              |
| <i>Subtask 1 — Question Reformulation</i>      |              |         |              |       |       |              |
|                                                | Score        | R1      | R2           | RLsum | BERT  | Align        |
| Dual + note grounding + pattern scoring        | <b>27.09</b> | 30.06   | 10.08        | 28.23 | 40.65 | 19.68        |
| <i>Subtask 2 — Evidence Identification</i>     |              |         |              |       |       |              |
|                                                | $\mu$ P      | $\mu$ R | $\mu$ F1     | mP    | mR    | mF1          |
| 10-shot, union + post-proc                     | 52.34        | 75.71   | 61.90        | 56.68 | 75.52 | <b>61.10</b> |
| <i>Subtask 3 — Answer Generation</i>           |              |         |              |       |       |              |
|                                                | Score        | BLEU    | RLsum        | SARI  | BERT  | Align        |
| o3 + GPT-5.2 + GPT-5.1                         | <b>30.95</b> | 9.11    | 23.13        | 56.54 | 37.24 | 22.07        |
| <i>Subtask 4 — Evidence Alignment</i>          |              |         |              |       |       |              |
|                                                | $\mu$ P      | $\mu$ R | $\mu$ F1     | mP    | mR    | mF1          |
| Ensemble + rescue heuristics                   | 83.30        | 77.70   | <b>80.41</b> | 84.83 | 80.72 | 81.81        |

Table 1: Results across subtasks with Development Phase experiments shown first and Test Phase results summarized at the end. <sup>†</sup> indicates configurations using explicit reasoning or chain-of-thought(CoT) prompting. <sup>‡</sup> indicates baseline configurations without additional grounding or pattern-scoring components.

## 7. Future Work

Future work will focus on improving data scale, modeling integration, and generation reliability. Larger clinical QA datasets would enable more reliable evaluation and tuning. Joint modeling across subtasks may reduce error propagation. Improving Subtask 3 is a priority, as answer generation remains the weakest component. The task requires producing short, faithful summaries grounded in sparse evidence. We will explore stronger grounding constraints and improved generation strategies. We will further investigate lightweight models that approximate ensemble performance at lower computational cost. This includes fine-tuning on larger EHR corpora and tighter integration between subtasks, such as using ST4 alignments to guide ST3 generation.

## 8. Limitations

This work has several limitations. First, the dataset is small. The development set contains only 20 cases. This limit restricts reliable hyperparameter tuning. It also limits the diversity of few-shot examples. Second, the system relies heavily on prompt engineering. Model behavior depends on prompt wording and example selection. Small prompt changes can alter predictions. This sensitivity reduces reproducibility across deployments. Third, the pipeline treats subtasks mostly independently. Errors can propagate between stages. For example, missing evidence in Subtask 2 may affect answer generation in Subtask 3. The current system does not perform joint reasoning across subtasks. Fourth, the approach depends on large proprietary models. These models require external APIs and significant computational cost. This constraint may limit reproducibility for researchers without access to similar infrastructure.

## 9. Ethics Statement

This work uses de-identified clinical data provided by the shared task. No identifiable patient information is used. The system is intended for research purposes only and not for clinical deployment.

## 10. Bibliographical References

- Tom B. Brown et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.
- Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. 2009. [What can natural language processing do for clinical decision support?](#) *Journal of Biomedical Informatics*, 42(5):760–772.
- Samah Fodeh, Linhai Ma, Ganesh Puthiaraju, Srivani Talakokkul, Afshan Khan, Ashley Haggaman, Sarah R. Lowe, and Aimee Kendall Roundtree. 2026. [TAB-PO: Preference optimization with a token-level adaptive barrier for token-critical structured generation.](#) *arXiv preprint arXiv:2603.00025*.
- Yao Fu, Hao Peng, and Ashish Sabharwal. 2023. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.
- Tianyu Gao et al. 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.
- Or Honovich et al. 2022. [TRUE: Re-evaluating factual consistency evaluation.](#) In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920. Association for Computational Linguistics.
- Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.
- Ziwei Ji, Tianyu Yu, Yifan Xu, Nayeon Lee, Eiichiro Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self-reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843. Association for Computational Linguistics.
- Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 540–551.
- Leonid Kuligin, Jan Lammert, Alexander Ostapenko, Kim Bressen, Martin Boeker, and Malte Tschochohei. 2025. [Prompt design for medical question answering with large language models.](#) *Machine Learning with Applications*, page 100758.
- Sunjun Kweon, Jiyoung Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwang Hyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. EHRNoteQA: An LLM benchmark for real-world clinical practice using discharge summaries. In *Advances in Neural Information*



- Processing Systems (Datasets and Benchmarks Track)*.
- Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C. Wallace. 2023. Do we still need clinical language models? *arXiv preprint arXiv:2302.08091*.
- 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. ArXiv:2005.11401.
- Hongfang Liu et al. 2013. An information extraction framework for cohort identification using electronic health records. *Journal of Biomedical Informatics*, 46(5):S35–S42.
- L. Liu et al. 2025. [Using natural language processing to extract information from clinical text for clinical registries](#). *Journal of the American Medical Informatics Association*.
- Harsha Nori et al. 2023. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.
- Rafael Rafailov et al. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
- Hannah Rashkin, Yonatan Bisk, and Yejin Choi. 2021. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3214–3252. Association for Computational Linguistics.
- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.
- Yash Saxena, Shubham Chopra, and A. M. Tripathi. 2024. [Evaluating consistency and reasoning capabilities of large language models](#). In *Second International Conference on Data Science and Information System (ICDSIS)*, pages 1–5. IEEE.
- Karan Singhal et al. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.
- Sarvesh Soni and Dina Demner-Fushman. 2026a. [A dataset for addressing patient’s information needs related to clinical course of hospitalization](#). *Scientific Data*.
- 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.
- Xuezhi Wang et al. 2023. Self-consistency improves chain-of-thought reasoning in language models. In *International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*. ArXiv:2201.11903.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 401–415.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.
- Denny Zhou, Nathanael Schärli, Le Hou, et al. 2023. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations (ICLR)*.



## 11. Appendix

This section presents the prompt templates used across all four subtasks as (Figure A1). Each template specifies the instruction structure, role configuration, and output format used during inference. The templates correspond to the prompting strategies described in section 4 and illustrate the constraints applied to each task. Subsections below describe the prompt design for Subtasks 1–4.

### Prompt Templates for Subtasks

#### Subtask 1: Clinician-Interpreted Question

**Role:** `user` only (Claude Sonnet 4 / GPT-4o). Preceded by clinical understanding pre-pass (extracts decisions, conditions, urgency, intent type) and  $N$  dev few-shot examples.

**Core instruction:** Generate 5 clinician-interpreted question candidates. Do NOT add severity/inference words (*low, high, elevated*) unless explicit in patient question. Preserve urgency indicators (*emergency, emergent, salvage*) before the procedure name. Abstract procedural details to clinical intent.

**Output format:** CANDIDATE\_1: [q]  
... CANDIDATE\_5: [q] ( $\leq 15$  words, ends ?, no first-person. Best selected by pattern-alignment scoring against dev gold templates.

#### Subtask 2: Evidence Selection

**Role:** `user` only (ensemble: o3 + GPT-5.2 + GPT-5.1). 10 dev few-shot examples (leave-one-out).

**Core instruction:** You are a clinical evidence selector. Given a patient question, a clinician-interpreted question, and numbered note sentences, output the *minimal* set of sentence IDs needed to answer the question.

**Merge:** union vote (`min_votes=1`) across 3 models; invalid IDs dropped post-hoc.

**Output format:** ["1", "3", "7"] (JSON array of strings; [] if none)

#### Subtask 3: Answer Generation (two-stage faithful)

**Role:** `user` only (cascade: o3→GPT-5.2→GPT-5.1). Using 10 dev few-shot examples.

**Stage 1** — cited draft: “You are a medical expert. Rephrase the relevant note sentences to answer the question. Stay close to the original wording. Cite each used sentence as [2], [7]. 70–75 words. Third person. Do NOT add interpretation beyond the note.”

**Stage 2** — faithfulness rewrite: “Rewrite the draft using ONLY the cited sentences above. No citation markers in output. Same medical terms and values. 70–75 words.”

#### Subtask 4: Evidence Alignment

**Role:** `system` + interleaved `user/assistant` few-shot (up to 20 dev cases, leave-one-out).

**System:** “You are a clinical evidence alignment expert. Include a note sentence only when it states the same specific fact, number, or event as the answer sentence (same lab value, dose, procedure, timeline). When in doubt, prefer inclusion (recall matters).”

**User turn** adds: `clinician_answer_without_citation` paragraph for full-answer context.

**Merge:** majority vote  $\theta = \lfloor MS/2 \rfloor + 1$ ;  $\theta$  swept on dev gold.

```
[{"answer_id": "1",
 "evidence_id": ["3", "7"]},
 {"answer_id": "2",
 "evidence_id": ["5"]},
 {"answer_id": "3",
 "evidence_id": []}]
```

# Razreshili at ArchEHR-QA 2026: Evidence Alignment via LLM Prompting and Cross-Encoder Fine-tuning

Arina Zemchik

arina.zemchik@gmail.com

## Abstract

We describe our system for Subtask 4 (Evidence Alignment) of the ArchEHR-QA 2026 shared task, which requires aligning each sentence of a clinician-authored answer to the supporting sentence(s) in a clinical note excerpt derived from MIMIC (Johnson et al., 2016). The task is challenging due to many-to-many alignment structure, answer sentences with no note support, and the semantic gap between clinical note language and answer paraphrases. We explore two approaches: few-shot chain-of-thought prompting with Qwen2.5-7B-Instruct and LoRA fine-tuning of a cross-encoder with combined InfoNCE and BCE loss. Our best system achieves a micro F1 of 67.93 on the test set.

**Keywords:** clinical NLP, evidence alignment, cross-encoder, parameter-efficient fine-tuning

## 1. Introduction

Grounded question answering over electronic health records requires not only generating accurate answers but also explicitly identifying where in the record each claim originates. The ArchEHR-QA 2026 shared task (Soni and Demner-Fushman, 2026b) formalises this challenge in four complementary subtasks. Subtask 4, which we address, requires aligning each sentence of a reference clinician-authored answer to the specific supporting sentence(s) in a clinical note excerpt, given the patient question, clinician-interpreted question, note excerpt, and reference answer as input.

The task presents several distinct challenges. First, alignment is many-to-many: a single note sentence may support multiple answer sentences, and a single answer sentence may draw on multiple note sentences. Second, answer sentences are paraphrases or aggregations of note content rather than direct extractions, requiring semantic rather than lexical matching. Third, approximately 19% of answer sentences in the development set are drawn from clinician domain knowledge rather than the note and should receive empty citation sets (Soni and Demner-Fushman, 2026a). Correctly identifying these cases is critical for precision.

We explore two complementary approaches. The first fine-tunes a cross-encoder using LoRA with a combined contrastive and calibration loss, leveraging human-annotated relevance labels from the dataset as a source of hard negatives. The second uses few-shot prompting of a quantized large language model with chain-of-thought examples that explicitly model the grounded versus inference-based distinction. Our experiments show that the LLM approach generalises better to unseen test cases, achieving micro F1 of 67.93 versus 64.44 for the fine-tuned cross-encoder.

## 2. Related Work

Evidence alignment is closely related to sentence-level evidence retrieval in extractive question answering and fact verification, where systems must identify minimal supporting spans for a given claim (Thorne et al., 2018; Rajpurkar et al., 2016). More recently, claim-level grounding has emerged as a distinct problem: frameworks such as eTracer (Chu et al., 2026) and AGREE (Ye et al., 2024) align individual response claims to supporting source sentences, treating grounding as a post-hoc verification step rather than retrieval. The ArchEHR-QA task differs in that alignment targets answer sentences rather than atomic claims, and some answer sentences have no note support - a distinction not present in standard extractive QA or citation generation benchmarks.

Cross-encoders have been widely used for passage relevance ranking due to their ability to jointly encode query-document pairs (Nogueira and Cho, 2019). Contrastive objectives such as InfoNCE (van den Oord et al., 2018) improve ranking quality but produce scores that are invariant to query-level shifts, making them poorly suited for threshold-based retrieval. Sheikholeslami et al. (2025) propose the Mann-Whitney (MW) loss, which directly maximises AUC by minimising binary cross-entropy over pairwise score differences, producing globally calibrated scores. Our combined InfoNCE and BCE loss addresses the same score compression problem through a simpler regularisation term, though MW loss represents a more principled future direction.

Parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022) and IA<sup>3</sup> (Liu et al., 2022) adapt pretrained models with few additional parameters, making them attractive for shared tasks with limited training data.

Large language models have demonstrated strong performance on clinical reasoning tasks

via few-shot prompting (Singhal et al., 2023), and chain-of-thought prompting improves multi-step reasoning (Wei et al., 2022). In the 2025 iteration of this shared task, the runner-up system (Bogireddy et al., 2025) used DSPy-optimised prompts with self-consistency voting for sentence-level evidence identification, outperforming fine-tuned retrieval approaches by a substantial margin - consistent with our finding that LLM prompting generalises better than cross-encoder fine-tuning in low-data clinical settings. Our prior work (Zemchyk, 2025) explored contrastive fine-tuning of a bi-encoder for retrieval-augmented answer generation in the same task family; here we extend this to evidence alignment with a cross-encoder and a combined ranking and calibration loss.

### 3. System Description

#### 3.1. Data and Task Setup

The development set contains 20 cases with 95 answer sentences, of which 77 have gold citations and 18 are inference-based with no supporting note sentences. Note sentences are annotated with relevance labels: *essential*, *supplementary*, and *not-relevant*. The test set contains 147 cases. Evaluation uses micro F1 as the primary metric.

#### 3.2. Approach 1: Zero-shot Cross-encoder Baseline

We evaluated four candidate cross-encoders on the development set: `ms-marco-MiniLM-L-6-v2` (general, fast), `ms-marco-MiniLM-L-12-v2` (general, larger), `pritamdeka/S-PubMedBert-MS-MARCO` (biomedical domain-adapted), and `ncbi/MedCPT-Cross-Encoder` (biomedical). Despite the biomedical models being domain-matched to the clinical task, `ms-marco-MiniLM-L-12-v2` achieved the highest development macro F1 after threshold tuning, consistent with prior findings that MS MARCO cross-encoders transfer strongly to out-of-domain retrieval (Thakur et al., 2021; Rosa et al., 2022). We therefore selected it as the base model for both the zero-shot baseline and subsequent fine-tuning.

For inference, we applied a two-stage strategy: if the maximum sigmoid score across all note sentences falls below a no-citation threshold  $t_{nc}$ , the answer sentence is assigned an empty citation set; otherwise, all note sentences scoring above a citation threshold  $t_c$  are returned, with a fallback to the top-scoring sentence. Thresholds were tuned on the development set ( $t_{nc} = 0.2$ ,  $t_c = 0.4$ ), achieving a development macro F1 of 0.723. Two implementation details were critical: applying sigmoid to raw logits before thresholding, and correctly scoring

empty-vs-empty citation pairs as  $F1 = 1.0$  rather than 0.0.

#### 3.3. Approach 2: Cross-Encoder Fine-tuning with Two-Stage Inference

We fine-tuned `ms-marco-MiniLM-L-12-v2` (Nguyen et al., 2016) using LoRA ( $r=4$ ,  $\alpha=8$ ) with a two-stage inference pipeline that separates the citation decision from evidence ranking, allowing each to be calibrated independently.

**Stage 1: Citation Decision.** For each answer sentence, we compute sigmoid scores across all note sentences. If the maximum score falls below a no-citation threshold  $t_{nc}$ , the sentence is predicted as inference-based and assigned an empty citation set.

**Stage 2: Evidence Ranking.** For sentences passing Stage 1, all note sentences scoring at or above a citation threshold  $t_c$  are cited, with a fallback to the single highest-scoring sentence. This stage is where fine-tuning contributes most directly - the LoRA-adapted model is trained to rank gold citation sentences above hard negatives.

**Training Data.** We constructed 138 contrastive groups from all 20 development cases. For each (answer sentence, gold citation) pair, hard negatives were sampled using a three-tier priority: (1) *supplementary* sentences, human-annotated as relevant but not essential - the most confusable negatives available; (2) sentences within a  $\pm 2$  position window of the gold citation; (3) other *not-relevant* sentences. Easy negatives (section headers, sentences under 6 words) were excluded. Inference-based answer sentences were excluded from training entirely, as including them consistently degraded the citation decision in Stage 1.

**Loss Function.** We trained with a combined loss:

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \lambda \cdot \mathcal{L}_{\text{BCE}} \quad (1)$$

where  $\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(s_+/ \tau)}{\sum_i \exp(s_i / \tau)}$  with temperature  $\tau = 0.2$ , and  $\mathcal{L}_{\text{BCE}}$  is binary cross-entropy with the positive labelled 1.0 and negatives 0.0, weighted by  $\lambda = 0.3$ . InfoNCE trains Stage 2 ranking, while BCE anchors absolute score magnitudes to prevent compression that would destabilise the Stage 1 threshold.

Models were trained for 8 epochs with a learning rate of  $2 \times 10^{-4}$ , linear warmup over 10% of steps, and AdamW optimiser with weight decay 0.01.

**Threshold Tuning.** Both thresholds were tuned jointly on the development set via grid search over  $t_{nc} \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$  and  $t_c \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ , optimising macro F1. The fine-tuned model required higher thresholds ( $t_{nc} = 0.6$ ,  $t_c = 0.7$ ) than the zero-shot model ( $t_{nc} = 0.2$ ,  $t_c = 0.4$ ), reflecting the upward shift in score magnitudes produced by the BCE calibration term.

An initial attempt used InfoNCE alone without sigmoid calibration, treating thresholds as operating on raw logits; this achieved micro F1 of 59.95 with high precision (77.16) but poor recall (49.02), confirming that uncalibrated scores make threshold tuning unreliable.

All experiments were run on a single T4 GPU.

### 3.4. Approach 3: LLM Few-shot Prompting

We used Qwen2.5-7B-Instruct (Qwen et al., 2025) running in 4-bit quantization via bitsandbytes on a single T4 GPU. Inference over the full test set of 147 cases required approximately 30 minutes, compared to under one minute for the fine-tuned cross-encoder including threshold tuning on the development set. The full prompt template is provided in Appendix A.

The prompt included three few-shot examples with chain-of-thought reasoning:

- a single-citation case
- a multi-citation case
- an inference-based case with empty citations

The third example was critical; without it the model consistently over-cited inference sentences. The prompt explicitly distinguished sentences that *directly state* the answer from sentences that are merely related.

Outputs were constrained to JSON format.

## 4. Results

Table 1 presents the performance of the evaluated systems on the ArchEHR-QA Subtask 4 test set. The Qwen-based system outperformed all cross-encoder variants. While the LoRA fine-tuned model substantially improved recall compared to the early cross-encoder baseline (65.66 vs. 49.02), this came at the cost of reduced precision. The Qwen approach achieved the best overall balance between precision and recall, resulting in the highest micro F1 score.

## 5. Discussion

### 5.1. LLM Prompting vs. Cross-Encoder Fine-tuning

Despite being fine-tuned on the task with human-annotated hard negatives, the LoRA cross-encoder (micro F1 64.44) underperformed the few-shot LLM (micro F1 67.93). We attribute this to two complementary factors: the nature of the alignment task itself, and the limitations imposed by the small development set.

Evidence alignment requires determining whether an answer sentence is *semantically entailed* by a note sentence - a reasoning operation that goes beyond surface similarity. Cross-encoders pre-trained on MS MARCO learn to match queries to relevant passages, but clinical entailment involves paraphrase, aggregation, and implicit inference that differ structurally from passage retrieval. Qwen2.5-7B, by contrast, can reason explicitly about whether a claim is *directly stated* in the note or represents clinician inference, as demonstrated by its ability to correctly handle inference-based sentences when provided with chain-of-thought examples.

From a practical standpoint, the cross-encoder approach offers a substantial efficiency advantage: the full pipeline including development set threshold tuning runs in under one minute on a T4 GPU, compared to approximately 30 minutes for LLM inference over the same number of cases. This motivates further work on closing the performance gap through better contrastive training objectives and calibration methods.

### 5.2. Representation Collapse Under Fine-tuning

A recurring failure mode across our fine-tuning experiments was score compression: after training, sigmoid scores for all note sentences clustered in a narrow range, shifting the optimal citation threshold from 0.4 (zero-shot) toward 0.3 or lower. In severe cases, the model collapsed entirely, predicting empty citations for all answer sentences (NoCiteAcc = 1.0, F1  $\approx$  0.35).<sup>1</sup>

We hypothesise that this reflects a form of representation collapse specific to small contrastive datasets. With only 138 training pairs across 20 cases, the model encounters the same note sentences repeatedly across different training examples. The LoRA layers adapt to reduce loss on these specific sentence pairs, but in doing so compress the representation space: note sentences

<sup>1</sup>NoCiteAcc is the proportion of inference-based answer sentences (those with empty gold citation sets) correctly predicted as having no supporting note sentence.



| System                               | Precision    | Recall       | F1           |
|--------------------------------------|--------------|--------------|--------------|
| Early cross-encoder (no sigmoid fix) | <b>77.16</b> | 49.02        | 59.95        |
| LoRA + combined loss                 | 63.27        | <b>65.66</b> | 64.44        |
| Qwen2.5-7B few-shot (best)           | 73.78        | 62.95        | <b>67.93</b> |

Table 1: Test set performance on ArchEHR-QA Subtask 4.

that were distinguishable under the pre-trained model become more similar in the fine-tuned embedding space, making threshold-based discrimination harder rather than easier. This is consistent with known collapse phenomena in contrastive learning (Jing et al., 2022), where insufficient diversity in training pairs leads to degenerate representations.

Adding a BCE calibration term to the InfoNCE loss partially mitigated this by anchoring absolute score magnitudes, improving dev macro F1 from 0.704 to 0.760. However, the large gap between dev macro F1 (0.760) and test micro F1 (64.44) suggests the calibration was itself overfit to the score distribution of the 20 dev cases rather than generalising to the 147 test cases.

### 5.3. Threshold Tuning as the Dominant Factor

Our two-stage inference design separates the citation decision (Stage 1) from evidence ranking (Stage 2), with fine-tuning primarily targeting the latter. However, our results suggest that threshold calibration for Stage 1 contributed more to overall performance than the ranking improvements from fine-tuning.

Cross-encoder with optimised thresholds ( $t_{nc} = 0.2$ ,  $t_c = 0.4$ ) achieved dev macro F1 of 0.723, comparable to or exceeding several fine-tuned variants. The first submitted model, which lacked sigmoid calibration and used untuned thresholds, scored only 59.95 micro F1 - well below what the zero-shot model with proper calibration would have achieved. Applying sigmoid transformation to raw logits and separately tuning  $t_{nc}$  and  $t_c$  on the development set was the single largest source of improvement across all experiments.

Fine-tuning improved Stage 2 ranking as intended (recall increased substantially from the first to final submission (49.02  $\rightarrow$  65.66)) but at the cost of precision (77.16  $\rightarrow$  63.27). The model learned to cite more aggressively in Stage 2 without learning when to abstain in Stage 1, suggesting that the ranking and abstention objectives interfere when trained jointly on a small dataset. The BCE term partially decoupled these by calibrating absolute scores, but the Stage 1 threshold still had to shift substantially ( $t_{nc} : 0.2 \rightarrow 0.6$ ) to compensate for the upward score drift introduced by training.

### 5.4. The Inference Sentence Problem

Correctly identifying inference-based answer sentences - those with empty citation sets - proved to be a surprisingly tractable problem for the zero-shot cross-encoder. At  $t_{nc} = 0.2$ , the model correctly identified 15 of 18 such sentences on the dev set (NoCiteAcc = 0.833) without any task-specific training, simply because these sentences have low semantic overlap with any note content. The three failures involved sentences that, while inference-based, happened to be topically related to note content (e.g., discussing a procedure mentioned in the note). Fine-tuning degraded this capability in most experiments, likely because excluding inference sentences from training left the model without signal about *when* to abstain from citing.

## 6. Conclusion

We presented two approaches for clinical evidence alignment: cross-encoder fine-tuning with a combined ranking and calibration loss, and LLM few-shot prompting.

The LLM approach achieved the best performance (micro F1 67.93), outperforming fine-tuned cross-encoders on a task that requires semantic reasoning over clinical text. However, the cross-encoder approach is approximately 30 $\times$  faster at inference, making it more suitable for deployment settings where low latency matters. Future work could explore hybrid approaches that retain the reasoning quality of LLMs while reducing inference cost, such as using LLM predictions as soft labels for cross-encoder distillation, or using cross-encoder scores as lightweight features within the LLM prompt.

## 7. Limitations

Results are based on a small development set of 20 cases, which limits the reliability of threshold tuning and ablation conclusions. Additionally, the Qwen model was run in 4-bit quantization, which may reduce performance compared to full precision.



## 8. Bibliographical References

- Sai Prasanna Teja Reddy Bogireddy, Abrar Ma-jeedi, Viswanath Gajjala, Zhuoyan Xu, Siddhant Rai, and Vaishnav Potlapalli. 2025. [Neural at ArchEHR-QA 2025: Agentic prompt optimization for evidence-grounded clinical question answering](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 104–109. Association for Computational Linguistics.
- Bohao Chu, Qianli Wang, Hendrik Damm, Hui Wang, Ula Muhabbek, Elisabeth Livingstone, Christoph M. Friedrich, and Norbert Fuhr. 2026. etracer: Towards traceable text generation via claim-level grounding. *arXiv preprint arXiv:2601.03669*.
- 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*.
- Li Jing, Pascal Vincent, Yann LeCun, and Yuan-dong Tian. 2022. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*.
- 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.
- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated MACHine reading COMprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*, volume 1773. CEUR-WS.org.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.
- 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. *arXiv preprint arXiv:2412.15115*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.
- Guilherme Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. In defense of cross-encoders for zero-shot retrieval. *arXiv preprint arXiv:2212.06121*.
- Nima Sheikholeslami et al. 2025. Optimizing what matters: AUC-driven learning for robust neural retrieval. *arXiv preprint arXiv:2510.00137*.
- 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*, 620:172–180.
- Sarvesh Soni and Dina Demner-Fushman. 2026a. [A dataset for addressing patient’s information needs related to clinical course of hospitalization](#). *Scientific Data*.
- 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.
- 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*.
- 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. 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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Xi Ye, Ruoxi Sun, Sercan Ö. Arik, and Tomas Pfister. 2024. [Effective large language model adaptation for improved grounding and citation generation](#). 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 6237–6251. Association for Computational Linguistics.

Arina Zemchyk. 2025. razreshili at ArchEHR-QA 2025: Contrastive fine-tuning for retrieval-augmented biomedical QA. In *Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP) at ACL 2025*.

## A. LLM Prompt Template

We design a structured prompt consisting of a system instruction and a user prompt with few-shot examples to guide the model in identifying evidence sentences.

### System Prompt

You are a clinical NLP expert. Your task is to identify which sentences from a clinical note directly support a given answer sentence.

You must respond with ONLY a JSON object – no explanation, no markdown, no extra text.

Follow these rules strictly:

- INCLUDE a sentence if it directly states the same information as the answer sentence (verbatim or paraphrased).
- INCLUDE all sentences that each contribute a distinct piece of information present

in the answer sentence. – DO NOT include a sentence just because it is adjacent to or on the same topic as a supporting sentence – it must independently state part of the answer. – DO NOT include a sentence that only provides context, background, or consequence of what the answer states. – Return [] if no note sentence directly states the answer content.

### User Prompt with Few-Shot Examples

```
-- [EXAMPLE 1 (single supporting sentence; enforces precision, avoids over-citation)]
--
-- [EXAMPLE 2 (multi-sentence candidate set; enforces selecting only directly supporting evidence, ignoring adjacent/context sentences)]
--
-- [EXAMPLE 3 (no supporting evidence; enforces empty output when answer is not grounded in the note)] --
-- END EXAMPLES --
```

Now do the same for:

Answer sentence:  
"<ANSWER\_SENTENCE>"

Clinical note sentences:  
<NOTE\_BLOCK>

Which note sentence IDs directly support the answer sentence?

Respond with ONLY this JSON:  
"evidence\_id": ["<id>", ...]

# Neural at ArchEHR-QA 2026: One Method Fits All: Unified Prompt Optimization for Clinical QA over EHRs

Abrar Majeedi<sup>1,\*</sup>, Viswanatha Reddy Gajjala<sup>1,\*</sup>,  
Sai Prasanna Teja Reddy Bogireddy<sup>2,\*</sup>, Siddhant Rai<sup>3,\*</sup>

<sup>1</sup>University of Wisconsin–Madison <sup>2</sup>University of Chicago <sup>3</sup>Independent Researcher  
{amajeedi, vgajjala}@wisc.edu, bogireddytejareddy@uchicago.edu

\*Equal contribution

## Abstract

Automated question answering (QA) over electronic health records (EHRs) demands precise evidence retrieval, faithful answer generation, and explicit grounding of answers in clinical notes. In this work, we present *Neural1.5*, our method for the ArchEHR-QA 2026 shared task at CL4Health@LREC 2026, which comprises of four subtasks: question interpretation, evidence identification, answer generation, and evidence alignment. Our approach decouples the task into independent, modular stages and employs DSPy’s MIPROv2 optimizer to automatically discover high-performing prompts, jointly tuning instructions and few-shot demonstrations for each stage. Within every stage, self-consistency voting over multiple stochastic inference runs suppresses spurious errors and improves reliability, while stage-specific verification mechanisms (e.g., self-reflection and chain-of-verification for alignment) further refine output quality. Among all teams that participated in all four subtasks, our method ranks second overall (mean rank 4.00), placing 4th, 1st, 4th, and 7th on Subtasks 1–4, respectively. These results demonstrate that systematic, per-stage prompt optimization combined with self-consistency mechanisms is a cost-effective alternative to model fine-tuning for multi-faceted clinical QA.

**Keywords:** clinical question answering, prompt optimization, electronic health records, large language models, evidence grounding, evidence alignment

## 1. Introduction

Patient medical advice requests have surged 55% since 2019, with physicians now spending 24% more time on inbox management (Arndt et al., 2024). Automatically answering these questions using electronic health records (EHRs) could substantially reduce clinician burden, yet requires systems that not only generate accurate responses but also explicitly ground every claim in verifiable clinical evidence. The ArchEHR-QA 2026 shared task at CL4Health@LREC 2026 (?) takes a step towards addressing this challenge through four complementary subtasks: (1) transforming verbose patient questions into concise clinician-interpreted queries, (2) identifying minimal evidence sentences from clinical notes, (3) generating grounded answers, and (4) aligning each answer sentence to its supporting evidence. Together, these subtasks capture the complete pipeline from question understanding to explainable, evidence-backed clinical QA.

The natural language processing capabilities of Large Language Models (LLMs) present a promising approach. Although LLMs have demonstrated strong performance in clinical QA (Singhal et al., 2025), their deployment faces two critical barriers. First, fine-tuning on clinical data is constrained by limited supervised datasets and overfitting risks. Second, prompt engineering for multi-stage clinical workflows remains labor-intensive and domain-specific, requiring expert iteration to identify opti-

mal instructions and demonstrations (Karayanni et al., 2024). Existing automated prompt optimization techniques (Wang et al., 2023) typically treat tasks holistically, failing to leverage the modular structure inherent in clinical QA pipelines where evidence identification, answer generation, and citation alignment each demand distinct reasoning patterns.

In this work, we present **Neural1.5**, a modular LLM method that addresses all four ArchEHR-QA 2026 subtasks through *automated, stage-specific prompt optimization*. Our key design principle is systematic task decomposition: by defining clear evaluation objectives for each subtask, we enable DSPy’s MIPROv2 optimizer (Khatab et al., 2024) to automatically discover high-performing prompts that jointly tune instructions and few-shot demonstrations. We further enhance reliability through *self-consistency voting* (Wang et al., 2022) for evidence identification and a *three-stage alignment pipeline* combining initial alignment, self-reflection to prune false positives, and chain-of-verification with confidence-weighted majority voting.

Our proposed method achieves competitive results across all four subtasks, ranking **4th** (Subtask 1), **1st** (Subtask 2 with Strict Micro F1 of 63.7), **4th** (Subtask 3), and **7th** (Subtask 4), with a **mean rank of 4.00** across all subtasks—**second overall** among teams participating in all four subtasks. These results demonstrate that systematic, per-stage prompt optimization combined with self-

consistency mechanisms offers a cost-effective alternative to model fine-tuning for multi-faceted clinical QA.

The key contributions of our work are:

- **Full-Pipeline Automated Optimization:** We propose the method to apply automated prompt optimization independently to all four ArchEHR-QA subtasks, with stage-specific objectives that capture distinct reasoning requirements for question interpretation, evidence retrieval, answer generation, and citation alignment.
- **Three-Stage Alignment with Self-Reflection:** We introduce an alignment pipeline that combines initial sentence-level alignment, self-reflection to identify and remove spurious citations, and chain-of-verification, with confidence-weighted majority voting across multiple stochastic runs to suppress hallucinations/errors.
- **Consistent Multi-Task Performance:** With a mean rank of 4.00 across all four subtasks, we demonstrate that modular, optimization-driven approaches can maintain strong performance across diverse clinical QA challenges, offering a practical blueprint for deploying LLMs in safety-critical healthcare applications.

## 2. Related Work

**Clinical QA:** Developing QA systems for clinical data has long been an interest in biomedical NLP. Earlier datasets like emrQA (Pampari et al., 2018) generated large-scale QA pairs from electronic medical records. Recent research has shown that LLMs can achieve near-expert performance on medical QA benchmarks (Singhal et al., 2025). The ArchEHR-QA shared task (Soni and Demner-Fushman, 2026b; Soni et al., 2025) advances this line of work by requiring systems to ground answers explicitly in clinical notes. Our prior work (Bogireddy et al., 2025) demonstrated the effectiveness of automated prompt optimization for a single task, achieving second place overall. Building on these insights, the 2026 edition expands to four complementary subtasks spanning question interpretation, evidence identification, answer generation, and evidence alignment.

**Prompt Optimization:** There is growing interest in automated prompt search. Methods such as APE (Zhou et al., 2022) and OPRO (Yang et al., 2023) treat prompt design as a black-box optimization problem. MIPRO (Opsahl-Ong et al., 2024) extends this to multi-stage LLM programs, jointly optimizing instructions and demonstration examples. Our work leverages MIPROv2 (Opsahl-Ong et al., 2024), which uses a combination of prompt proposal and Bayesian search.

**Self-Consistency:** LLMs can produce variable outputs given the same prompt. The self-consistency decoding strategy (Wang et al., 2022) addresses this by sampling multiple outputs and selecting the most consistent result.

## 3. Task Description

The ArchEHR-QA 2026 shared task (Soni and Demner-Fushman, 2026b) comprises four subtasks. The dataset (Soni and Demner-Fushman, 2026a) consists of patient-authored questions, clinician-interpreted counterparts, clinical note excerpts with sentence-level relevance annotations, and reference clinician-authored answers with answer–evidence alignments.

**Subtask 1: Question Interpretation.** Given a free-text patient question, generate a concise clinician-interpreted question ( $\leq 15$  words) that captures the core clinical information need.

**Subtask 2: Evidence Identification.** Given the patient question, clinician question, and a clinical note excerpt with numbered sentences, identify the minimal set of note sentences that provide evidence needed to answer the question.

**Subtask 3: Answer Generation.** Given the questions and clinical note excerpt, generate a grounded natural-language answer ( $\leq 75$  words) using only information from the notes.

**Subtask 4: Evidence Alignment.** Given the questions, clinical note excerpt, and a reference answer with numbered sentences, align each answer sentence to its supporting note sentence(s).

## 4. Methodology

Our method draws on a human-inspired decoupling strategy, separating question understanding, evidence gathering, answer formulation, and evidence attribution into distinct stages. We operationalize this intuition as a modular pipeline, with each subtask addressed by a DSPy program whose prompts are optimized independently. The initial prompt templates (DSPy signatures) for all subtasks are provided in Appendix A.

### 4.1. Subtask 1: Question Interpretation

We define a DSPy *Signature* for question interpretation that instructs the LLM to transform a patient narrative into a concise clinician-interpreted question. The signature encodes domain-specific

heuristics: preserving key medical terms exactly (procedure names, medication names, condition names), using patient-specific phrasing (“him/her/the patient”), and following high-scoring question patterns (e.g., “Why was [X] recommended to him/her?”).

The module uses `ChainOfThought` prompting and enforces a strict 15-word limit through post-processing. We optimize the prompt using MIPROv2 with a composite metric combining semantic entailment (via an LLM-as-judge evaluator assessing AlignScore-like semantic equivalence), key term preservation, and structural conformity.

**Prompt-Optimization Objective:** MIPROv2 searches the space of instructions and few-shot exemplars to maximize a weighted composite of semantic alignment (60%), key term preservation (25%), and question structure (15%) on the development set.

## 4.2. Subtask 2: Evidence Identification

For each question–note pair, we classify each note sentence as essential or irrelevant. Given the clinical note with sentences  $s_1, s_2, \dots, s_n$  and gold labels  $y_i \in \{\text{essential}, \text{supplementary}, \text{not-relevant}\}$ , the model predicts binary labels  $\hat{y}_i \in \{0, 1\}$ .

The classification uses a two-step process. First, for each sentence, we generate reasoning about why it is or is not essential using dedicated reasoning signatures (`EssentialReasoning` and `NonEssentialReasoning`). These reasoning traces serve as few-shot demonstrations for the main classifier. Second, a `MedicalAnswerWithCitations` signature classifies all sentences jointly, providing a relevancy score (0–10) and reasoning for each.

**Prompt-Optimization Objective:** We invoke MIPROv2 to optimize the classification prompt, maximizing the sentence-level  $F_1$  between predicted and gold essential sentences:

$$F_1(Y^+, \hat{Y}^+), \quad Y^+ = \{i \mid y_i = \text{essential}\}, \\ \hat{Y}^+ = \{i \mid \hat{y}_i = 1\}.$$

**Self-Consistency Voting:** The classifier is executed  $R = 5$  times on the same input with stochastic sampling (temperature 0.8). The final label is obtained by majority vote:

$$v_i = \sum_{r=1}^R \hat{y}_i^{(r)}, \quad \hat{y}_i = \begin{cases} 1 & \text{if } v_i \geq \lceil R/2 \rceil, \\ 0 & \text{otherwise.} \end{cases}$$

This suppresses spurious single-run errors and retains sentences identified as essential by at least three of five passes.

## 4.3. Subtask 3: Answer Generation

Given a question  $q$  and the set of essential sentences  $E = \{s_i \mid \hat{y}_i = 1\}$  from Subtask 2, Stage 3 produces a concise answer  $a_{\text{gen}} (\leq 75 \text{ words})$ . The `GroundedMedicalAnswer` signature encodes key constraints: the answer must use only facts from the clinical notes, preserve exact clinical terminology, write in professional register, and avoid citation markers.

To further improve answer quality, we employ a consolidation step: the model generates  $R = 5$  candidate answers using stochastic sampling (temperature 0.9), and a separate consolidation prompt selects claims consistently supported across candidates to produce a single final answer.

**Prompt-Optimization Objective:** MIPROv2 optimizes the answer generation prompt using an LLM-as-judge metric that evaluates: faithfulness to the clinical notes, medical completeness (concept coverage), lexical similarity to the reference, and coherence (structure and register).

## 4.4. Subtask 4: Evidence Alignment

Given a question  $q$ , clinical note sentences, and reference answer sentences  $a_1, \dots, a_m$ , the method must produce alignment links  $(a_j, s_i)$  mapping each answer sentence to its supporting note sentence(s).

We introduce a **three-stage pipeline**:

**Stage A – Initial Alignment:** The `EvidenceAlignment` signature instructs the model to align each answer sentence to note sentences that directly support it, outputting confidence scores for each link. The prompt explicitly warns against both over-citing and under-citing.

**Stage B – Self-Reflection:** The `SelfReflection` signature critically reviews the initial alignment to identify and remove false positive links (indirect or inferential connections) and add any clearly missing links. This step primarily targets precision improvement.

**Stage C – Chain-of-Verification:** The `ChainOfVerification` signature performs a final verification pass, checking each link against three criteria: (i) does the answer sentence directly reference information from the note sentence? (ii) would removing the note sentence cause the answer sentence to lose specific evidence? (iii) is the connection direct rather than inferential?

**Confidence-Weighted Majority Voting:** The full three-stage pipeline is executed  $R$  times with stochastic sampling. For each answer–evidence



link, we count votes across runs and compute average confidence. A link is retained only if it receives at least  $\lceil R/2 \rceil$  votes *and* the average confidence exceeds a threshold  $\tau_c = 0.9$ . This dual filtering mechanism balances recall and precision.

## 5. Experimental Setup

**Dataset:** We evaluated our method on the ArchEHR-QA 2026 dataset (Soni and Demner-Fushman, 2026a), which contains patient questions alongside clinical note excerpts derived from the MIMIC database, with sentence-level relevance annotations and reference answers with answer–evidence alignments. The development set comprises 20 cases (IDs 1–20) used for prompt optimization. Test set sizes vary by subtask: Subtasks 1–3 evaluate on 47 cases (IDs 121–167), while Subtask 4 evaluates on 147 cases (IDs 21–167), reflecting the staged data release schedule.

**Evaluation Metrics:** Each subtask is evaluated independently. **Subtask 1** (Question Interpretation) is evaluated using ROUGE, BERTScore, AlignScore, and MEDCON. **Subtask 2** (Evidence Identification) uses Precision, Recall, and F1 over predicted vs. gold evidence sentences, with strict (essential-only) and lenient (including supplementary) variants. **Subtask 3** (Answer Generation) uses BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), SARI (Xu et al., 2016), BERTScore (Zhang et al., 2020), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023). **Subtask 4** (Evidence Alignment) uses Precision, Recall, and F1 over predicted alignment links.

**LLM Configuration:** Across all subtasks, we employ the GPT-4.1 model accessed via the Azure OpenAI API. We chose GPT-4.1 for its strong instruction-following capabilities and large context window, which are critical for processing lengthy clinical notes. For prompt optimization with MIPROv2, we use temperature 0.3 to ensure deterministic optimizer feedback. For self-consistency runs, we use temperature 0.7 (Subtask 2), 0.8 (Subtask 4), and 0.9 (Subtask 3) to capture model uncertainty. The maximum context window is set to 10,000 tokens for Subtasks 2–4 and 2,000 tokens for Subtask 1. All other decoding parameters (e.g., top-p, frequency and presence penalties) are held at API defaults. We estimate the computational cost per case at approximately 5 LLM calls for Subtask 1, 25 calls (5 self-consistency runs  $\times$  5 reasoning steps) for Subtask 2, 6 calls for Subtask 3 (5 candidates + 1 consolidation), and 15–25 calls for Subtask 4 (3 stages  $\times$   $R$  runs).

| Team                    | Overall     | R.L.        | B.S.        | A.S.        | M.C.        |
|-------------------------|-------------|-------------|-------------|-------------|-------------|
| HealthNLLP_Ret.         | <b>31.2</b> | <b>35.3</b> | <b>46.8</b> | <b>24.0</b> | 18.7        |
| KPSCMI                  | <u>30.8</u> | <u>27.8</u> | <u>41.0</u> | <u>26.4</u> | <b>27.9</b> |
| OptiMed                 | 29.9        | 28.8        | 43.1        | 27.7        | 19.9        |
| <b>Neural1.5 (Ours)</b> | 28.9        | 31.3        | 43.6        | 15.2        | <u>25.6</u> |
| Yale-DM-Lab             | 27.1        | 28.2        | 40.6        | 19.7        | 19.8        |

Table 1: Subtask 1: Question Interpretation results on the test set. **Bold** = best, underlined = second best.

## 6. Results

Tables 1–4 present per-subtask results, and Table 5 summarizes rankings across all four subtasks.

**Subtask 1: Question Interpretation.** Our method achieves an overall score of 28.9, ranking 4th among 13 teams (Table 1). Our MEDCON score of 25.6 is the second highest, indicating strong preservation of medical concepts. The relatively lower AlignScore (15.2) suggests room for improvement in semantic alignment with reference questions, possibly due to differences in question framing style.

**Subtask 2: Evidence Identification.** Our method ranks **1st** with an overall Strict Micro F1 of 63.7 (Table 2), demonstrating the effectiveness of the prompt-optimized classification with self-consistency voting. Notably, our method achieves strong performance across all evaluation granularities: Strict Micro (P=60.2, R=67.6, F1=63.7), Lenient Micro (P=77.8, R=67.6, F1=72.4), and the highest Lenient Macro F1 of 73.1. This balanced profile avoids the extreme precision-recall trade-offs seen in some competing methods—for example, Yale-DM-Lab achieves the highest recall (75.7) but at the cost of the lowest precision (52.3).

**Subtask 3: Answer Generation.** We rank 4th with an overall score of 35.2 (Table 3). Our method achieves competitive BLEU (9.4) and AlignScore (34.3) results, with the latter being the second highest among all teams. Our MEDCON score of 40.9 ranks 4th, indicating reasonable preservation of medical concepts in the generated answers. The answer consolidation step, which aggregates five candidate answers via majority-supported claims, helps ensure factual consistency.

**Subtask 4: Evidence Alignment.** Our method achieves a Micro F1 of 78.6, ranking 7th among 15 teams (Table 4). The three-stage pipeline (alignment, self-reflection, chain-of-verification) combined with confidence-weighted majority voting yields high precision (84.3), though recall (73.7)

| Team                    | Overall     | Strict Micro |             |             | Lenient Micro |             |             | Strict Macro |             |             | Lenient Macro |             |             |
|-------------------------|-------------|--------------|-------------|-------------|---------------|-------------|-------------|--------------|-------------|-------------|---------------|-------------|-------------|
|                         |             | P            | R           | F1          | P             | R           | F1          | P            | R           | F1          | P             | R           | F1          |
| <b>Neural1.5 (Ours)</b> | <b>63.7</b> | <b>60.2</b>  | <b>67.6</b> | <b>63.7</b> | <b>77.8</b>   | <b>67.6</b> | <b>72.4</b> | <b>64.3</b>  | <b>69.7</b> | <b>64.8</b> | <b>81.9</b>   | <b>69.7</b> | <b>73.1</b> |
| OptiMed                 | 63.2        | 56.7         | 71.3        | 63.2        | 73.1          | 71.3        | 72.2        | 61.5         | 73.0        | 64.1        | 78.1          | 73.0        | 72.9        |
| UIC-AIHealth4All        | 62.9        | 59.3         | 67.0        | 62.9        | 74.3          | 67.0        | 70.4        | 61.8         | 69.2        | 63.0        | 77.8          | 69.2        | 70.5        |
| Yale-DM-Lab             | 61.9        | 52.3         | <b>75.7</b> | 61.9        | 68.0          | <b>75.7</b> | 71.6        | 56.7         | <b>75.5</b> | 61.1        | 73.8          | <b>75.5</b> | 70.4        |
| TJU222                  | 61.4        | 54.6         | 70.0        | 61.4        | 74.4          | 70.0        | <b>72.2</b> | 61.4         | 70.3        | 60.6        | 78.9          | 70.3        | 70.7        |
| TAMU-NLP-Lab            | 60.9        | 56.7         | 65.9        | 60.9        | 76.0          | 65.9        | 70.6        | 63.0         | 66.7        | 59.9        | 80.0          | 66.7        | 69.1        |

Table 2: Subtask 2: Evidence Identification results on the test set.

| Team                    | Overall     | BLEU       | R.L.        | SARI        | B.S.        | A.S.        | M.C.        |
|-------------------------|-------------|------------|-------------|-------------|-------------|-------------|-------------|
| WisPerMed               | <b>36.3</b> | <b>9.9</b> | <b>27.8</b> | 58.6        | <b>46.8</b> | <b>31.7</b> | <b>43.1</b> |
| TAMU-NLP-Lab            | 36.2        | 9.7        | 27.4        | 59.6        | 46.2        | 34.5        | 39.9        |
| BIT.UA-AAUBS            | 35.6        | 8.6        | 26.4        | <b>60.0</b> | 45.0        | 30.2        | 43.2        |
| <b>Neural1.5 (Ours)</b> | 35.2        | 9.4        | 25.6        | 57.7        | 43.5        | 34.3        | 40.9        |
| HealthNLP_Ret.          | 34.6        | 7.0        | 25.4        | 59.2        | 43.8        | 33.6        | 38.7        |
| OptiMed                 | 34.5        | 5.7        | 25.2        | 56.5        | 43.1        | 37.4        | 39.1        |

Table 3: Subtask 3: Answer Generation results on the test set.

| Team                    | Overall     | M.P.        | M.R.        | M.F1        |
|-------------------------|-------------|-------------|-------------|-------------|
| BIT.UA-AAUBS            | <b>81.5</b> | <b>88.0</b> | 75.9        | <b>81.5</b> |
| WisPerMed               | 81.3        | 86.9        | 76.3        | 81.3        |
| Yale-DM-Lab             | 80.4        | 83.3        | 77.7        | 80.4        |
| OptiMed                 | 80.3        | 80.7        | <b>79.8</b> | 80.3        |
| UIC-AIHealth4All        | 79.8        | 83.6        | 76.3        | 79.8        |
| tt501                   | 79.1        | 78.2        | 80.1        | 79.1        |
| <b>Neural1.5 (Ours)</b> | 78.6        | 84.3        | 73.7        | 78.6        |
| KPSCMI                  | 78.1        | 86.2        | 71.5        | 78.1        |

Table 4: Subtask 4: Evidence Alignment results on the test set. Overall = Micro F1.

| Team                    | ST1      | ST2      | ST3      | ST4      | Mean↓       |
|-------------------------|----------|----------|----------|----------|-------------|
| OptiMed                 | 3        | 2        | 6        | 4        | 3.75        |
| <b>Neural1.5 (Ours)</b> | 4        | <b>1</b> | 4        | 7        | <b>4.00</b> |
| WisPerMed               | 6        | 12       | <b>1</b> | 2        | 5.25        |
| HealthNLP_Ret.          | <b>1</b> | 7        | 5        | 9        | 5.50        |
| KPSCMI                  | 2        | 8        | —        | 8        | 6.00        |
| Yale-DM-Lab             | 5        | 4        | 13       | 3        | 6.25        |
| TAMU-NLP-Lab            | 11       | 6        | 2        | —        | 6.33        |
| BIT.UA-AAUBS            | 13       | 10       | 3        | <b>1</b> | 6.75        |

Table 5: Summary of rankings across all four subtasks for teams that participated in at least three subtasks. Mean Rank is computed over participated subtasks (lower is better). “—” = non-participation.

is somewhat lower. The conservative confidence threshold ( $\tau_c = 0.9$ ) favors precision over recall, reflecting our design choice to minimize false alignment links.

**Cross-Subtask Consistency.** Table 5 highlights a key strength of our approach: *consistent perfor-*

*mance across all four subtasks.* With a mean rank of 4.00 across subtasks, our method is second only to OptiMed (3.75) among teams participating in all four subtasks. While several teams achieve top ranks on individual subtasks (e.g., WisPerMed on ST3, BIT.UA-AAUBS on ST4), they exhibit more variance across the full task suite. Our modular prompt optimization framework delivers robust performance without specializing in any single subtask at the expense of others.

## 7. Conclusion

We present a modular approach for all four subtasks of the ArchEHR-QA 2026 shared task, leveraging DSPy’s MIPROv2 optimizer to autonomously discover high-performing prompts for each stage. For Subtask 1, the method transforms patient narratives into concise clinician queries optimized for semantic alignment. For Subtask 2, sentence-level evidence classification with self-consistency voting achieves the best F1 score among all participants. For Subtask 3, answer generation with multi-candidate consolidation produces grounded, clinically faithful responses. For Subtask 4, a novel three-stage alignment pipeline with self-reflection and chain-of-verification enables precise answer-evidence grounding. Across all four subtasks, our method achieves a mean rank of 4.00 (Table 5), the second best among teams participating in all subtasks, underscoring the consistency of our modular design.

Our results demonstrate that systematic prompt optimization, combined with self-consistency mechanisms, is a cost-effective and competitive alternative to model fine-tuning across diverse clinical QA

tasks. Future work may explore integrating external medical knowledge, cross-subtask feedback (e.g., using Subtask 1 outputs to improve Subtask 2), and extending the approach to longer clinical documents.

## 8. Limitations

Despite competitive performance, our method has several limitations. The pipeline treats each subtask largely independently, missing potential synergies (e.g., using evidence identification results to constrain answer generation). The self-consistency mechanism increases computational cost by a factor of  $R$  (typically 3–5 runs per input). The confidence threshold for Subtask 4 was tuned on the small development set (20 cases) and may not generalize optimally. Additionally, the method relies on GPT-4.1, making it dependent on a proprietary API, and has not been evaluated on notes from institutions beyond MIMIC. The conservative design choices (high confidence thresholds, strict majority voting) favor precision over recall, which may not be ideal for all clinical use cases.

## 9. Prompts and Code Availability

To promote transparency and reproducibility, we release all manual and optimized prompt templates, together with our full pipeline implementation at our GitHub repository.<sup>1</sup> The initial prompt templates for all subtasks are included in Appendix A.

## 10. Bibliographical References

- Brian G Arndt, Mark A Micek, Adam Rule, Christina M Shafer, Jeffrey J Baltus, and Christine A Sinsky. 2024. [More tethered to the EHR: EHR workload trends among academic primary care physicians, 2019–2023](#). *Annals of Family Medicine*, 22(1):12–18.
- Sai Prasanna Teja Reddy Bogireddy, Abrar Majeedi, Viswanath Gajjala, Zhuoyan Xu, Siddhant Rai, and Vaishnav Potlapalli. 2025. Neural at archehr-qa 2025: Agentic prompt optimization for evidence-grounded clinical question answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 104–109.
- Nader Karayanni, Aya Awwad, Chein-Lien Hsiao, and Surish P Shanmugam. 2024. Keeping experts in the loop: Expert-guided optimization for clinical data classification using large language models. *arXiv preprint arXiv:2412.02173*.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines. In *Proceedings of the Twelfth International Conference on Learning Representations*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. *arXiv preprint arXiv:2406.11695*.
- Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.
- 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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.
- Sarvesh Soni and Dina Demner-Fushman. 2026a. [A dataset for addressing patient's information needs related to clinical course of hospitalization](#). *Scientific Data*.
- 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.
- 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.

<sup>1</sup><https://github.com/bogireddytejareddy/ArchEHR-QA-Neural>

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P Xing, and Zhiting Hu. 2023. Promptagent: Strategic planning with language models enables expert-level prompt optimization. *arXiv preprint arXiv:2310.16427*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Wen-wai Yim, Yajuan 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. *arXiv preprint arXiv:2305.16739*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#).

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

## A. Prompt Templates

This appendix presents the core prompt templates (DSPy signatures) used in our method. These are the *initial* templates provided to MIPROv2; the optimizer refines the instructions and selects few-shot demonstrations automatically. Full optimized prompts are available in our repository.

### A.1. Subtask 1: Question Interpretation

#### Prompt Template: Question Interpretation

Transform a patient’s narrative into a concise clinical question ( $\leq 15$  words) that a clinician would need to answer by reviewing the patient’s medical record.

#### Core Constraints:

1.  $\leq 15$  words, strictly enforced.
2. Patient-specific: use “him/her/the patient” — never generic.
3. Preserve medical terms: use exact procedure/medication names from the narrative.
4. Must end with a question mark.

#### High-Scoring Patterns:

| Patient’s concern             | Target pattern                                         |
|-------------------------------|--------------------------------------------------------|
| “Why did they do X?”          | “Why was [X] recommended to him/her?”                  |
| “Will I recover?”             | “What is the expected course of recovery for him/her?” |
| “Why X instead of Y?”         | “Why was [X] recommended over [Y]?”                    |
| “Why was I given medication?” | “Why was he/she given [medication]?”                   |
| “Is this related to...?”      | “Are his/her [symptoms] related to [condition]?”       |

**Input:** Patient narrative.

**Output:** Concise clinician question ( $\leq 15$  words).

*Inference mode:* ChainOfThought prompting with post-processing to enforce the word limit.

### A.2. Subtask 2: Evidence Identification

#### Prompt Template: Reasoning Demonstrations

**Essential Reasoning:** Given a patient narrative, patient question, clinician question, and a clinical note sentence, provide reasoning for why this note sentence *is essential* to address the question.

**Non-Essential Reasoning:** Given the same inputs, provide reasoning for why this note sentence *is not essential* to address the question.

*Purpose:* These two prompts are used at training time to generate per-sentence reasoning traces, which serve as few-shot demonstrations for the classifier below.

#### Prompt Template: Evidence Classifier

You are a medical assistant.

1. You are provided with a patient narrative, patient question, and clinician question.
2. You are provided with the clinical notes related to the case.

Classify each clinical note sentence as either **essential** or **irrelevant** in addressing the patient question and clinician question. Provide a relevancy score (0–10) and reasoning for each.

Be **very critical** when assigning the essential tag. Only assign it if the note sentence is directly relevant to the specific question asked.

**First-Order Relevance Only:** A note is essential *only* if it directly answers or provides evidence for the question. Background context or treatment summaries are *not* essential.

**Input:** Patient narrative, patient question, clinician question, clinical notes.

**Output per note:** `<id>: <sentence> -> essential|irrelevant -> <score> -> <reasoning>`

**Inference:** Run  $R=5$  times at temperature 0.8; majority vote determines the final label.

### A.3. Subtask 3: Answer Generation

#### Prompt Template: Grounded Answer Generation

You are a medical assistant answering a patient's question using **only** information from the clinical note excerpt.

**Constraints:**

1. Answer must be at most 75 words (~5 sentences).
2. Use only facts stated in the clinical note. Do not add outside medical knowledge, generic advice, or speculation.
3. Write in professional clinical register (not simplified lay language).
4. Do not include citation markers such as [1], [2].
5. Reuse exact clinical wording and terminology from note sentences as much as possible.
6. The last sentence must directly answer the patient's question.

**Input:** Patient narrative, patient question, clinician question, clinical note excerpt.

**Output:** Concise grounded answer ( $\leq 75$  words).

**Inference:** Generate  $R=5$  candidates at temperature 0.9; consolidate via a separate prompt that retains only claims consistently supported across candidates.

#### Prompt Template: Answer Consolidation

You are a clinical answer consolidation system. Given a patient question and 5 candidate answers generated from a clinical note:

1. Retain only claims consistently supported across the candidate answers.
2. Ground strictly in clinical note content—do not add external knowledge or speculate.
3. Use professional medical register.
4. Limit to 75 words (~5 sentences).
5. Do not include patient names or identifying information.

Output only the final consolidated answer.

### A.4. Subtask 4: Evidence Alignment

#### Prompt Template: Stage A — Initial Alignment

You are a medical evidence alignment specialist. Align each answer sentence to the specific clinical note sentence(s) that **directly** support it.

**Alignment Rules:**

1. Align only when the answer sentence directly paraphrases, summarizes, or references information explicitly stated in the note sentence.
2. Do not align based on indirect associations, background context, or inferential connections.
3. Over-citing (unnecessary links) and under-citing (missing links) are both penalized.
4. Each answer sentence must be attributed to at least one note sentence. If no direct support exists, choose the closest note sentence and assign a low confidence (0.10–0.30).

**Input:** Patient narrative, patient question, clinician question, clinical note sentences, answer sentences.

**Output per answer sentence:**

`answer_sentence_k: [note_ids] (confidence=[scores])`

#### Prompt Template: Stage B — Self-Reflection

You are a strict reviewer performing self-reflection on an evidence alignment task. Critically review the initial alignment and identify:

1. **False positives** (primary focus): links where the answer does not directly use information from the linked note sentence. **Remove** these.
2. **False negatives** (secondary focus): missing links where an answer sentence clearly paraphrases or references a note sentence. **Add** only when direct and explicit.

Produce a corrected alignment with updated confidence scores.

**Additional input:** Initial alignment from Stage A.



**Prompt Template: Stage C — Chain-of-Verification**

You are a verification specialist. For **each** alignment link (answer sentence  $k \rightarrow$  note sentence  $i$ ), verify:

1. Does answer sentence  $k$  directly paraphrase or reference specific information from note sentence  $i$ ?
2. If note sentence  $i$  were removed, would answer sentence  $k$  lose a specific piece of evidence it relies on?
3. Is the connection direct (not through inference or intermediate reasoning)?

If **any** check fails, remove the link. Return the final verified alignment.

**Additional input:** Reflected alignment from Stage B.

*Post-hoc:* Run the full three-stage pipeline  $R$  times; retain a link only if votes  $\geq \lceil R/2 \rceil$  and average confidence  $\geq 0.9$ .

# OptiMed at ArchEHR-QA 2026: GEPA Prompt Optimization and Multi-Agent Majority Voting for EHR-Grounded Question Answering

Feras AlMannaa<sup>1</sup>, Talia Tseriotou<sup>2</sup>, Maria Liakata<sup>2,3</sup>

<sup>1</sup>Independent Researcher <sup>2</sup>Queen Mary University of London <sup>3</sup>The Alan Turing Institute  
{t.tseriotou, m.liakata}@qmul.ac.uk

## Abstract

Despite the demonstrated promise of Large Language Models in medical question answering, existing work largely addresses closed-form, exam-style tasks and overlooks complex open-ended questions requiring reasoning over noisy, long clinical documents. In this work, we present our system, **OptiMed**, submitted to the ArchEHR-QA 2026 shared task on grounded clinical question answering over EHR notes. We combine GEPA, an evolutionary prompt optimization framework, with multi-agent majority voting across five diverse LLMs and a structured clinical abstraction strategy for question interpretation. OptiMed ranked 1st overall among teams completing all four subtasks with an average score of 52.0, achieving top AlignScore in both Question Interpretation and Answer Generation, reflecting strong factual grounding. GEPA optimization proved effective for structured tasks with sufficient development data, but failed to generalize on complex generative tasks under very limited number of supervisions. Multi-agent majority voting consistently lifted performance in evidence-oriented subtasks. Prompt analysis attributes GEPA's gains to role prompting and procedural decomposition and failures to over-specification under limited supervision.

**Keywords:** Clinical Question Answering, Electronic Health Records, Prompt Optimization, Multi-Agent Framework

## 1. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in Medical Question Answering (QA), achieving expert-level performance across academic medical knowledge benchmarks and real-world clinical judgment tasks (Nori et al., 2023; Liévin et al., 2024; Singhal et al., 2025). Despite their success, such evaluations remain confined to exam-style, closed-form QA, failing to capture the complexity of open-ended patient questions that require reasoning over noisy, lengthy clinical documents (Bardhan et al., 2024) and grounding in the patient's unique medical history.

Yet few works have addressed the effectiveness of LLMs on Clinical QA over patient-specific records (Ahsan et al., 2024; AlMannaa et al., 2025), with most efforts focusing on condition identification or implicit diagnosis over EHRs (Pan et al., 2025; Albassam et al., 2025). Such tasks are well-structured and limited in scope. Other work focuses on EHR QA dataset development, which is instrumental in enabling downstream applications (Bardhan et al., 2022; Kweon et al., 2024).

With the growing burden of patient portal messaging on clinician time (Martinez et al., 2024; Mandal et al., 2024), LLM-powered QA workflows offer great promise for assisting clinicians through automatically generating EHR-grounded answers. Yet most efforts focus on patient question generation (Liu et al., 2024), with limited work on answer generation and existing studies of deployments relying on shallow, category-triggered EHR context with narrow evaluation of clinical adoption (Garcia

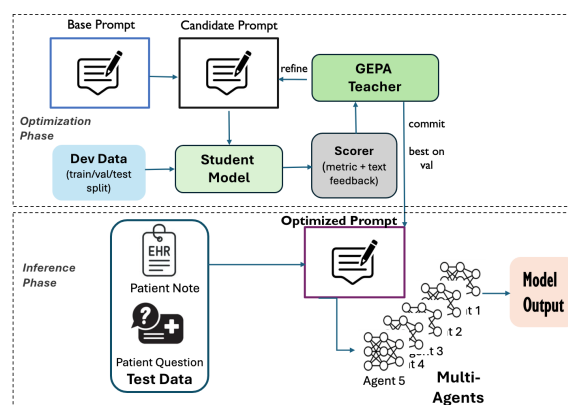


Figure 1: Overall Approach Diagram.

et al., 2024; Bootsma-Robroeks et al., 2025). The ArchEHR-QA shared task aims to bridge this gap through the development of benchmarks and systems for holistic patient QA, spanning from question interpretation to evidence-grounded answer generation over free-form EHR notes (Soni et al., 2025; Soni and Demner-Fushman, 2026).

While fine-tuning LLMs on EHR records offers a promising avenue, its usability is hindered by high computational costs, task diversity and the scarcity of quality clinical annotations. Prompting frameworks have emerged as a superior alternative, surpassing fine-tuned LLMs in both general Medical QA (Maharjan et al., 2024) and Clinical QA tasks (Ray et al., 2026). Yet the effectiveness of prompting is contingent on prompt quality. Prior work has explored prompt engineering including

few-shot and Chain of Thought (CoT) for Clinical NLP tasks (Sivarajkumar et al., 2024), yet despite impressive zero-shot performance, such prompts remain task-specific, hand-crafted and unreliable across diverse question types (Zhu et al., 2024). Such ad-hoc approaches lack scalability and generalizability for evidence retrieval (Karayanni et al., 2024), while recent work has shown that jointly optimizing instructions and few-shot demonstrations outperforms standard prompting on Clinical EHR QA (Bogireddy et al., 2025).

In this work, we use GEPA (Agrawal et al., 2025), an evolutionary prompt optimization framework, combining it with multi-agent majority-voting enabling faithful and well-grounded question interpretation and answer synthesis without manual prompt engineering. We demonstrate our approach in Figure 1.

## 2. Related Work

**Patient Question Interpretation** has been studied as consumer health question summarization (Ben Abacha and Demner-Fushman, 2019; Basu et al., 2025), with work exploring transfer (Yadav et al., 2021b) and reinforcement learning (Yadav et al., 2021a). Element-aware generation, where key medical entities are extracted before synthesis, reduces hallucinations and achieves the highest entailment without few-shot prompting (Basu et al., 2025), while larger LLMs in zero-shot settings surpass fine-tuned models and show lower error rate than experts (Van Veen et al., 2023). Our approach adopts an entity-aware, template-driven abstraction strategy directly motivated by these findings.

**Clinical EHR QA:** Despite being explored across structured and unstructured data, Clinical EHR QA still remains underexplored and challenging (Bardhan et al., 2024). While emrQA (Pampari et al., 2018) established QA grounded on clinical EHRs, EHRNoteQA (Kweon et al., 2024) extended this to real longitudinal notes. AIMannaa et al. (2025) show performance benefits of hybrid RAG pipelines for retrieving relevant EHR sections in open-form QA, while highlighting challenges with long context and longitudinal reasoning. McInerney et al. (2023) and Ahsan et al. (2024) explored zero-shot prompting for EHR feature extraction and evidence retrieval respectively, flagging hallucinations. With real-world deployments remaining shallow and lacking grounded evidence retrieval (Bootsma-Robroeks et al., 2025), ArchEHR-QA (Soni and Demner-Fushman, 2026) directly addresses evidence identification and alignment over clinical notes.

**Prompt Optimization** demonstrates consistent gains over manual designs on QA tasks (Yang et al., 2023; Guo et al., 2023). Karayanni et al. (2024)

show that expert-guided refinement outperforms both manual and automated optimization for clinical note classification, while Bogireddy et al. (2025) show that jointly optimizing instructions and demonstrations via MIPROv2 substantially outperforms zero-shot and few-shot baselines on grounded clinical note QA. Unlike MIPROv2’s Bayesian search over fixed proposals, GEPA samples task trajectories and reflects on them to diagnose prompt failures and evolve candidates, significantly outperforming MIPROv2 (Agrawal et al., 2025). Here we employ GEPA alongside a multi-agent framework.

**Multi-agent Voting:** Multi-agent frameworks including debating reduce hallucination and improve performance (Du et al., 2024; Mao et al., 2025), with majority voting identified as the primary driver of these gains (Choi et al., 2025). In medical settings, collaborative multi-agent frameworks show strong zero-shot performance (Kim et al., 2024; Tang et al., 2024), with Hwang et al. (2025) achieving top performance on grounded EHR QA via sequential agent collaboration with dynamic expert recruitment, and majority voting outperforming debating in clinical QA abstract screening (Akinseloyin et al., 2026). Building on these findings, we employ majority voting across five diverse LLMs to leverage complementary model strengths.

## 3. Task Description

The ArchEHR-QA shared task addresses patient health question answering grounded in EHRs, spanning four subtasks:

**T1 – Question Interpretation:** Given patient question  $q$ , generate a concise clinician-interpreted query  $\hat{q}'$  capturing the clinical rationale without introducing new facts.

**T2 – Evidence Identification:** Given  $q$ , clinician-interpreted query  $\hat{q}'$  and the numbered sentences from the clinical note excerpt,  $\{s_1, \dots, s_n\}$ , identify the *minimal sufficient set* of sentences  $Y^+ = \{i \mid \hat{y}_i = \text{essential}\}$  to answer  $q$ , remaining faithful to the excerpt even when it only partially addresses the question.

**T3 – Answer Generation.** Given  $q$ ,  $\hat{q}'$  and  $\{s_1, \dots, s_n\}$ , generate a concise grounded answer directly addressing  $q$ , without speculation or use of external knowledge.

**T4 – Evidence Alignment.** Given  $q$ ,  $\hat{q}'$ ,  $\{s_1, \dots, s_n\}$  and answer  $a = \{a_{s_1}, \dots, a_{s_m}\}$ , produce a many-to-many alignment mapping each answer sentence  $a_{s_j}$  to the note sentences that directly support it, yielding predicted pairs  $[a_{s_j}, s_i]$ .

### 3.1. Dataset and Evaluation

**Data:** The ArchEHR-QA 2026 dataset (Soni and Demner-Fushman, 2026) pairs patient-authored

questions  $q$  with clinical note excerpts from MIMIC (Johnson et al., 2016). Gold  $q'$  and  $a$  are provided where applicable (see Table 1). For Task 2 the sentence-level labels are: *essential* / *supplementary* / *non-relevant*.

| Subtask         | Dev | Test | Provided Inputs             |
|-----------------|-----|------|-----------------------------|
| T1: Patient Qn. | 120 | 47   | $q$                         |
| T2: Evidence    | 20  | 47   | $q, q', \{s_i\}_{i=1}^n$    |
| T3: Answer      | 20  | 47   | $q, q', \{s_i\}_{i=1}^n$    |
| T4: Alignment   | 20  | 147  | $q, q', \{s_i\}_{i=1}^n, a$ |

Table 1: Dataset splits and inputs per subtask.

**Evaluation:** Each subtask is evaluated with automatic metrics as shown in Table 2. For Task 2, *strict* (only *essential*) and *lenient* (not penalizing *supplementary*) F1 variants are reported.

| Metric                         | T1 | T2 | T3 | T4 |
|--------------------------------|----|----|----|----|
| P / R / F1                     |    | •  |    | •  |
| BLEU (Papineni et al., 2002)   |    |    | •  |    |
| ROUGE (Lin, 2004)              | •  |    | •  |    |
| SARI (Xu et al., 2016)         |    |    | •  |    |
| BERTScore (Zhang et al., 2019) | •  |    | •  |    |
| AlignScore (Zha et al., 2023)  | •  |    | •  |    |
| MEDCON (Yim et al., 2023)      | •  |    | •  |    |

Table 2: Evaluation metrics per subtask.

## 4. Methodology

Our approach implements a systematic multi-stage pipeline using DSPy (Khattab et al., 2023), applied independently per subtask. We first iterate over multiple DSPy signatures, selecting the best strategy per task (§4.1). Following model selection across various LLMs under standardized clinical constraints (§4.2), we apply GEPA prompt optimization (§4.3) with task-specific textual feedback signals. For structured prediction (Tasks 2 and 4), we further employ multi-agent voting to improve reliability (§4.4).

### 4.1. DSPy Signature Iteration

We define multiple DSPy signatures per subtask, iterating over single and multi-step modules across different task-solving approaches. We experiment with two prompting strategies, vanilla zero-shot and chain-of-thought (CoT), holding the base signature fixed during GEPA optimization.

### 4.2. Model and Strategy Selection

We evaluate several state-of-the-art LLMs optimized for reasoning and zero-shot capabilities, comparing performance across standard inference

and native reasoning modes. This ensures a standardized comparison of each model’s inherent ability to adhere to clinical constraints, helping us identify the best candidate for evidence-grounding and answer generation over EHRs.

### 4.3. GEPA Prompt Optimization

After establishing base performance, we apply GEPA (Agrawal et al., 2025) within DSPy to automatically optimize module instructions. GEPA operates as a teacher–student loop where a teacher model samples trajectories from the student model to analyze failures and propose improved instructions. Formally, GEPA searches for the optimal prompt  $P^*$ :

$$P^* = \arg \max_{P \in \mathcal{P}} \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(x, y^*) \in \mathcal{D}_{\text{train}}} \mathcal{R}(f_{\theta}(x; P), y^*) \quad (1)$$

where  $P \in \mathcal{P}$  is a candidate prompt,  $f_{\theta}(\cdot; P)$  is the student LLM conditioned on  $P$ ,  $(x, y^*)$  is a training pair and  $\mathcal{R}$  is the task-specific feedback reward.

Optimization targets two components sequentially:

- (1) **instruction tuning**, where the teacher diagnoses student errors and rewrites instructions evolved from the Pareto frontier and
- (2) **few-shot bootstrapping**, where high-quality demonstrations are automatically selected to accompany the optimized prompt.

We evaluate both stages during development. However, bootstrapped demonstrations did not overperform instruction-optimized zero-shot prompting, possibly due to the small number of examples available in dev set for demonstrations. Therefore all of our reported results reflect the instruction-optimized zero-shot prompting approach.

**Feedback Signal:** During initial evaluation (§4.1, 4.2) we use per-task metric averages (§3.1). During GEPA optimization, we additionally provide textual feedback alongside the numerical score to guide the teacher model when proposing new prompt candidates. For generative tasks (Tasks 1 and 3), an LLM judge (§A.2) generates a natural language explanation of how the student’s output diverges from the gold reference. For structured prediction tasks (Tasks 2 and 4), we construct explicit feedback that identifies false positives (incorrectly predicted sentences) and false negatives (missed sentences) in the predicted evidence or alignment lists.

### 4.4. Multi-Agent Voting

For structured Tasks 2 and 4, we aggregate predictions from five LLMs via voting. After testing

| Team                 | Avg         | T1          |             |             |             |             | T2          | T3          |            |             |             |             |             |             | T4          |
|----------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                      |             | O.          | R           | BS          | AS          | MED         | S.F1        | O           | B          | R           | SARI        | BS          | AS          | MED         | F1          |
| <b>OptiMed(ours)</b> | <b>52.0</b> | 29.9        | 28.8        | 43.1        | <b>27.7</b> | 19.9        | 63.2        | 34.5        | 5.7        | 25.2        | 56.5        | 43.1        | <b>37.4</b> | 39.1        | 79.1        |
| Neural               | 51.6        | 28.9        | 31.3        | 43.6        | 15.2        | 25.6        | <b>63.7</b> | 35.2        | 9.4        | 25.6        | 57.7        | 43.5        | 34.3        | 40.9        | 80.3        |
| WisPerMed            | 50.8        | 26.9        | 21.8        | 38.0        | 21.7        | <b>26.3</b> | 58.8        | <b>36.3</b> | <b>9.9</b> | <b>27.8</b> | 58.6        | <b>46.8</b> | 31.7        | 43.1        | <b>81.5</b> |
| HealthNLP_R.         | 50.7        | <b>31.2</b> | <b>35.3</b> | <b>46.8</b> | 24.0        | 18.7        | 60.2        | 34.6        | 7.0        | 25.4        | 59.2        | 43.8        | 33.6        | 38.7        | 79.8        |
| Yale-DM-Lab          | 50.1        | 27.1        | 28.2        | 40.6        | 19.7        | 19.8        | 61.9        | 31.0        | 9.1        | 23.1        | 56.5        | 37.2        | 22.1        | 37.6        | 67.2        |
| BIT.UA-AAUBS         | 48.7        | 19.0        | 17.9        | 29.4        | 8.7         | 20.2        | 58.8        | 35.6        | 8.6        | 26.4        | <b>60.0</b> | 45.0        | 30.2        | <b>43.2</b> | 80.4        |
| sebis                | 45.9        | 25.6        | 22.9        | 36.9        | 21.4        | 21.3        | 51.6        | 31.5        | 3.8        | 22.0        | 56.6        | 39.3        | 33.9        | 33.4        | 74.8        |

Table 3: Results on the ArchEHR-QA 2026 test set for teams completing all four subtasks, ranked by average overall score. O. = Overall, B = BLEU, R = ROUGELsum, BS = BERTScore, AS = AlignScore, MED = MEDCON (UMLS), S.F1 = Strict Micro F1, F1 = Micro F1.

different voting thresholds, we find strict majority ( $\geq 3/5$ ) to perform best:

$$\hat{y}_i = \mathbf{1} \left[ \sum_{k=1}^5 \hat{y}_i^{(k)} \geq 3 \right] \quad (2)$$

improving reliability and recall while suppressing single-model errors. The shared agent instruction was optimized via GEPA. We further investigate the effect of upstream evidence quality on Tasks 3 and 4 by using the multi-agent ensemble evidence from Task 2 as input in a two-step approach.

## 5. Experimental Setup

All models are accessed via OpenRouter with no fine-tuning.

**GEPA Teacher:** `gemini-3.0-flash`, selected for its performance, context size, and efficiency as the optimizer.

**Single-model Inference:** `glm-4.7`, with strong instruction-following and cost efficiency.

**Multi-agent Voting:** Ensemble of the five following zero-shot LLMs: `deepseek-v3.2`, `glm-4.7`, `gemini-3.0-flash`, `gpt-5.2`, and `kimi-k2.5`, maximizing diversity across training data, architecture, and provider, identified as the key driver of majority voting gains (Akinseloyin et al., 2026).

**GEPA Prompt Optimization Setup:** For GEPA optimization, the development data is partitioned into train, validation and held-out test subsets: 80/20/20 for Task 1 (120 examples) and 7/3/10 for Tasks 2, 3, and 4 (20 examples). GEPA iterates over the training subset, using textual feedback to propose new prompt candidates, committing each only if it improves performance on the validation subset. The held-out test subset is used for selecting the best-performing overall system configuration across models and approaches prior to final evaluation on the shared task test set.

**Inference Parameters:** All single-model runs use greedy decoding (temperature = 0), while ensemble runs use temperature = 0.3 to promote output

diversity. A maximum of 32,000 output tokens is used throughout. Where supported, native reasoning was disabled, as enabling it yielded worse performance on the dev set.

## 6. Results

Table 3 shows performance across teams completing all four subtasks. OptiMed ranked **1st** overall among 7 full-task teams with a score of **52.0**, and among all 23 teams ranked 3rd in T1 (**29.9**), 2nd in T2 (**63.2**), 6th in T3 (**34.5**), and 4th in T4 (**80.3**). Notably, we achieve the top AlignScore across all teams in both Tasks 1 and 3, reflecting superior factual grounding relative to the ground truth, avoiding hallucination which is particularly meaningful in medical settings. These results demonstrate that GEPA optimization and multi-agent majority voting complement each other across tasks, forming a strong candidate for clinical settings, though GEPA’s gains are contingent on sufficient dev set size for reliable prompt optimization. We discuss task-specific findings below.

| Model                         | Task 1      | Task 2      | Task 3      |
|-------------------------------|-------------|-------------|-------------|
| <code>glm-4.7</code>          | <b>35.7</b> | <b>50.1</b> | 37.3        |
| <code>gemini-3.0-flash</code> | 26.5        | 37.5        | <b>40.5</b> |
| <code>kimi-k2.5</code>        | –           | <b>50.1</b> | 37.1        |
| <code>gpt-5.2</code>          | 21.6        | 49.1        | 38.4        |
| <code>sonnet-4.5</code>       | 23.9        | 37.3        | –           |

Table 4: Model performance on dev set.

**Model Selection:** Table 4 shows individual model performance on the dev set (see Table 1). `glm-4.7` achieves the highest scores on Tasks 1 and 2, making it a strong choice for a single base model on these tasks. On Task 1, it specifically surpasses all other models by a very large margin. However, on Task 3 `gemini-3.0-flash` has the best performance, motivating its selection as the base model for answer generation using GEPA, with `glm-4.7` retained as a strong candidate for its consistently high performance across tasks.



| System            | O.          | R           | BS          | AS          | MED         |
|-------------------|-------------|-------------|-------------|-------------|-------------|
| Base prompt + CoT | 27.8        | 26.0        | 40.5        | 22.6        | <b>22.0</b> |
| GEPA              | 29.3        | 28.5        | 42.8        | 23.6        | 22.3        |
| GEPA + CoT        | <b>29.9</b> | <b>28.8</b> | <b>43.1</b> | <b>27.7</b> | 19.9        |

Table 5: Zero-shot performance for Task 1 with `glm-4.7` model on test set.

**Question Interpretation:** Motivated by the superiority of element-aware generation (Basu et al., 2025), our base prompt identifies the patient’s concern, extracts relevant medical entities (symptoms, events and therapies) and maps them to a standard query template based on the question type (treatment, cause, plan questions, care questions) with emphasis on clinical formalization. Clinical formalization refers to the structured reframing of informal patient language into a concise, clinician-facing query (see Appendix A.1). Table 6 shows results for three systems using `glm-4.7`, where GEPA yields a 1.5-point gain over the base prompt, with CoT providing further improvements.

| System                           | S.F1        | S.P         | S.R         | L.F1        | L.P         | L.R         |
|----------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| GEPA w. $q, q'$                  | 59.4        | <b>57.8</b> | 61.0        | 67.1        | <b>74.6</b> | 61.0        |
| GEPA w. $q$                      | 61.9        | 52.0        | <b>76.6</b> | 71.0        | 66.2        | <b>76.6</b> |
| GEPA w. $q$<br>+ Multi-agent (3) | <b>63.2</b> | 56.7        | 71.3        | <b>72.2</b> | 73.1        | 71.3        |

Table 6: Zero-shot performance on Task 2 with `glm-4.7` on test set on micro F1. S. = Strict, L. = Lenient.

**Evidence Identification:** With `glm-4.7` remaining the best on the dev set, we compare two GEPA variants: providing  $q$  only versus both  $q$  and  $q'$ . The latter yields higher Precision but lower Recall and overall lower F1, showing that  $q'$ ’s conciseness focuses evidence selection more narrowly. Multi-agent majority voting (3/5) across five diverse LLMs achieves the best overall performance.

| System                                       | O.          | B   | R           | SARI        | BS          | AS          | MED         |
|----------------------------------------------|-------------|-----|-------------|-------------|-------------|-------------|-------------|
| GEPA (Gemini)<br>Multi-Agent from T2 (thr=2) | 30.0        | 4.2 | 23.4        | 54.6        | 41.5        | 23.8        | 32.5        |
| GEPA (Gemini)<br>Multi-Agent from T2 (thr=3) | 31.1        | 5.2 | 23.9        | 55.0        | 42.1        | 26.3        | 33.9        |
| Base prompt (GLM)<br>Only GLM from T2        | <b>34.5</b> | 5.7 | <b>25.2</b> | <b>56.5</b> | <b>43.1</b> | <b>37.4</b> | <b>39.1</b> |

Table 7: Zero-shot performance on Task 3 test set. T2 denotes using the Task 2 evidence output as input in a *two-step* approach.

**Answer Generation:** We adopt a two-stage approach, using Task 2 evidence as input for Task 3. For generation, the simpler base prompt using `glm-4.7` significantly outperforms the GEPA variant using `gemini-3.0-flash` for both the majority (3 out of 5 LLMs) and lenient (2 out of 5 LLMs) multi-agent voting variants. Additionally, the majority (strict) variant of voting surpasses the lenient across all the metrics. We attribute GEPA’s failure to the task complexity and diversity in combination

with the small dev set, which provides insufficient signal to reliably distinguish prompt quality from instance-level variance.

| System                                           | Micro F1    | Micro P     | Micro R     |
|--------------------------------------------------|-------------|-------------|-------------|
| Base prompt + CoT<br>Multi-Agent from T2 (thr=3) | 62.6        | 64.3        | 61.0        |
| Base prompt + Multi-agent (3)                    | <b>80.3</b> | 80.7        | <b>79.8</b> |
| GEPA + Multi-agent (3)                           | 79.3        | <b>80.5</b> | 78.1        |

Table 8: Zero-shot performance on Task 4 with `glm-4.7` model on test set. T2 denotes using the Task 2 evidence output as input in a *two-step* approach, while Multi-agent refers to a *single-step* multi-agent approach without Task 2 input.

**Evidence Alignment:** Based on our Task 2 findings, we omit  $q'$ . A two-step approach using Task 2 evidence identification followed by `glm-4.7` CoT alignment yielded poor performance, likely due to error propagation and single-model high uncertainty. Switching to single-step multi-agent majority voting that does not use Task 2 as input improved results, with the base variant outperforming GEPA, consistent with our Task 3 finding that GEPA fails to generalize on small dev sets and complex tasks such as many-to-many evidence alignment.

## 6.1. GEPA-Optimized Prompts Insights

Base and GEPA-optimized prompts per task are provided in Appendix A.1. For Tasks 1 and 2, where GEPA outperformed the base prompt, gains appear driven by role prompting that places the model in a clinical context (i.e. *You are a reliable expert clinician*) and fine-grained procedural instructions that decompose the task into verifiable steps. For Tasks 3 and 4, the GEPA prompts are markedly more complex, introducing formal auditing frameworks and domain-specific sub-rules. We hypothesize that the combination of inherently more complex tasks with very limited training examples produces high variance in the optimization signal, causing the teacher model to over-specify instructions rather than learning robust task-level patterns, resulting in prompts that do not generalize well.

## 7. Conclusion

We present OptiMed, combining GEPA prompt optimization and multi-agent majority voting for grounded clinical EHR question answering, ranking 1st overall among teams completing all four subtasks and achieving the top AlignScore in generative tasks. GEPA proved effective where training signal was sufficient, while multi-agent voting yielded notable gains on complex structured prediction tasks. For complex tasks with small dev sets, GEPA underperformed base prompts, highlighting the importance of sufficient supervision signal.

## 8. Limitations

While GEPA optimization shows benefits, it is sensitive to development set size, conditioning its success on the quality and quantity of clinical annotations, especially for complex generative tasks. The real performance gains under greater data availability therefore remain to be examined. Regarding cost, multi-agent voting across five LLMs amplifies inference costs compared to single-LLM counterparts, which may hinder deployment in real clinical settings. Our system is evaluated on a single benchmark dataset, and thus generalization to other EHR systems with different structures and more diverse patient populations remains untested. Additionally, the effectiveness of such a system for patients from diverse racial and ethnic backgrounds remains unknown, as their conditions and clinical reasoning patterns may be underrepresented in LLM training data, potentially compromising both accuracy and fairness. Finally, while our system achieves strong automatic metric performance, the clinical validity of generated answers and evidence alignments has not been verified through human expert evaluation which is essential before any real-world deployment.

## 9. Ethics

This work uses the ArchEHR-QA 2026 dataset, which is derived from MIMIC (Johnson et al., 2016) and accessed under the PhysioNet Credentialed Health Data License 1.5.0. The dataset contains de-identified patient records in accordance with HIPAA Safe Harbor standards. All experiments were conducted solely on this de-identified data, and no attempts were made to re-identify any individuals.

## 10. Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible AI UK (KP0016) as a Keystone project lead by Maria Liakata.

## 11. Bibliographical References

- Lakshya A Agrawal, Shangyin Tan, Dilara Soyulu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. 2025. Gepa: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*.
- Hiba Ahsan, Denis Jered McInerney, Jisoo Kim, Christopher Potter, Geoffrey Young, Silvio Amir, and Byron C Wallace. 2024. Retrieving evidence from ehRs with llms: possibilities and challenges. *Proceedings of machine learning research*, 248:489.
- Opeoluwa Akinseloyin, Xiaorui Jiang, and Vasile Palade. 2026. Llm-based multi-agent collaboration for abstract screening towards automated systematic reviews. *Biology Methods and Protocols*, page bpag006.
- Dina Albassam, Adam Cross, and Chengxiang Zhai. 2025. Leveraging llms for predicting unknown diagnoses from clinical notes. *arXiv preprint arXiv:2503.22092*.
- Feras AlMannaa, Talia Tseriotou, Jenny Chim, and Maria Liakata. 2025. Investigating LLM capabilities on long context comprehension for clinical question answering. *arXiv preprint arXiv:2510.18691*.
- Jayetri Bardhan, Kirk Roberts, and Daisy Zhe Wang. 2024. Question answering for electronic health records: Scoping review of datasets and models. *J Med Internet Res*, 26:e53636.
- Abhishek Basu, Deepak Gupta, Dina Demner-Fushman, and Shweta Yadav. 2025. A dataset and benchmark for consumer health-care question summarization. *arXiv preprint arXiv:2512.23637*.
- Asma Ben Abacha and Dina Demner-Fushman. 2019. 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.
- Sai Prasanna Teja Reddy Bogireddy, Abrar Ma-jeedi, Viswanath Gajjala, Zhuoyan Xu, Siddhant Rai, and Vaishnav Potlapalli. 2025. Neural at ArchEHR-QA 2025: Agentic prompt optimization for evidence-grounded clinical question answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 104–109, Vienna, Austria. Association for Computational Linguistics.
- Charlotte M. H. H. T. Bootsma-Robroeks, Jessica D. Workum, Stephanie C. E. Schuit, Anne Hoekman, Tarannom Mehri, Job N. Doornberg, Tom P. van der Laan, and Rosanne C. Schoonbeek. 2025. Ai-generated draft replies to patient messages: exploring effects of implementation. *Frontiers in Digital Health*, Volume 7 - 2025.
- Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. 2025. Debate or vote: Which yields better decisions

- in multi-agent large language models? In *Advances in Neural Information Processing Systems*.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.
- 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–e243201.
- Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2023. Evoprompt: Connecting llms with evolutionary algorithms yields powerful prompt optimizers. *arXiv e-prints*, pages arXiv–2309.
- Hyeon Hwang, Hyeongsoon Hwang, Jongmyung Jung, Jaehoon Yun, Minju Song, Yein Park, Dain Kim, Taewhoo Lee, Jiwoong Sohn, Chanwoong Yoon, Sihyeon Park, Jiwoo Lee, Heechul Yang, and Jaewoo Kang. 2025. [DMIS lab at ArchEHR-QA 2025: Evidence-grounded answer generation for EHR-based QA via a multi-agent framework](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 118–125, Vienna, Austria. Association for Computational Linguistics.
- Nader Karayanni, Aya Awwad, Chein-Lien Hsiao, and Surish P Shanmugam. 2024. Keeping experts in the loop: Expert-guided optimization for clinical data classification using large language models. *arXiv preprint arXiv:2412.02173*.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: an adaptive collaboration of llms for medical decision-making. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.
- 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.
- Siru Liu, Aileen P Wright, Allison B McCoy, Sean S Huang, Julian Z Genkins, Josh F Peterson, Yaa A Kumah-Crystal, William Martinez, Babatunde Carew, Dara Mize, Bryan Steitz, and Adam Wright. 2024. [Using large language model to guide patients to create efficient and comprehensive clinical care message](#). *Journal of the American Medical Informatics Association*, 31(8):1665–1670.
- Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. [Can large language models reason about medical questions?](#) *Patterns*, 5(3):100943.
- Jenish Maharjan, Anurag Garikipati, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Rahul Thapa, Qingqing Mao, and Ritankar Das. 2024. Openmedlm: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, 14(1):14156.
- Soumik Mandal, Batia M Wiesenfeld, Devin M Mann, Adam C Szerencsy, Eduardo Iturrate, and Oded Nov. 2024. Quantifying the impact of telemedicine and patient medical advice request messages on physicians' work-outside-work. *NPJ Digital Medicine*, 7(1):35.
- Junyu Mao, Anthony Hills, Talia Tseriotou, Maria Liakata, Aya Shamir, Dan Sayda, Dana Atzil-Slonim, Natalie Djohari, Arpan Mandal, Silke Roth, et al. 2025. [Automated data enrichment using confidence-aware fine-grained debate among open-source LLMs for mental health and online safety](#). *arXiv preprint arXiv:2512.06227*.
- Kathryn A Martinez, Rebecca Schulte, Michael B Rothberg, Maria Charmaine Tang, and Elizabeth R Pfoh. 2024. Patient portal message volume and time spent on the ehr: an observational study of primary care clinicians: Martinez et al. *Journal of General Internal Medicine*, 39(4):566–572.
- Denis McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron Wallace. 2023. [CHILL: Zero-shot custom interpretable feature extraction from clinical notes with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8477–8494, Singapore. 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*.
- Jie Pan, Seungwon Lee, Cheliger Cheliger, Elliot A. Martin, Kiarash Riazi, Hude Quan, and Na Li. 2025. [Integrating large language models with human expertise for disease detection in electronic health records](#). *Computers in Biology and Medicine*, 191:110161.
- 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.
- Sushant Kumar Ray, Gautam Siddharth Kashyap, Sahil Tripathi, Nipun Joshi, Vijay Govindarajan, Rafiq Ali, Jiechao Gao, and Usman Naseem. 2026. Do clinical question answering systems really need specialised medical fine tuning? *arXiv preprint arXiv:2601.12812*.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950.
- 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 Med Inform*, 12:e55318.
- Sarvesh Soni and Dina Demner-Fushman. 2026. 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.
- 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, Viena, Austria. Association for Computational Linguistics.
- 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.
- Dave Van Veen, Cara Van Uden, Louis Blanke-meier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. 2023. Clinical text summarization: adapting large language models can outperform human experts. *Research square*, pages rs–3.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. [Optimizing statistical machine translation for text simplification](#). volume 4, pages 401–415.
- Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2021a. [Reinforcement learning for abstractive question summarization with question-aware semantic rewards](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 249–255, Online. Association for Computational Linguistics.
- Shweta Yadav, Mourad Sarrouiti, and Deepak Gupta. 2021b. [NLM at MEDIQA 2021: Transfer learning-based approaches for consumer question and multi-answer summarization](#). In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 291–301, Online. Association for Computational Linguistics.
- Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.
- 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. *arXiv preprint arXiv:1904.09675*.

Yinghao Zhu, Zixiang Wang, Junyi Gao, Yuning Tong, Jingkun An, Weibin Liao, Ewen M Harrison, Liantao Ma, and Chengwei Pan. 2024. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint arXiv:2402.01713*.

## 12. Language Resource References

Jayetri Bardhan, Anthony Colas, Kirk Roberts, and Daisy Zhe Wang. 2022. [DrugEHRQA: A question answering dataset on structured and unstructured electronic health records for medicine related queries](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1083–1097, Marseille, France. European Language Resources Association.

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.

Sunjun Kweon, Jiyoung Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwang Kim, Jee-won 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.

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.

Sarvesh Soni and Dina Demner-Fushman. 2026. [A dataset for addressing patient’s information needs related to clinical course of hospitalization](#). *Scientific Data*.

## A. Appendix

### A.1. Base versus GEPA-Optimized Prompts

For each Task we show the hand-crafted **Base** prompt alongside the **GEPA-optimized** prompt.

## Task 1 – Question Interpretation

### Base Prompt – Task 1

Analyze a patient’s story to figure out the patient’s main concern/question and formulate the single question a clinician would use to look up in the EHR record.

#### Task:

1. Locate the specific *Conflict*, *Gap of information*, or *Confusion*:
  - Is it about a Treatment Choice?
  - Is it about a Cause?
  - Is it about the Future/Plan?
  - Is it about Quality of Care?
2. Identify the main *Event/Condition*, *Symptom*, and *Therapy*:
  - Are any of these directly related to the patient’s question?
  - If yes, only then include them in the output question.
3. Map to a Standard Clinical Query Template:
  - Treatment → “Why was [Therapy] recommended/given?”
  - Cause → “Is [Symptom] related to [Event/Condition]?”
  - Plan → “What is the expected course of recovery?” or “What should the patient do?”
  - Care → “Why was [Diagnosis/Treatment] delayed?” or “Why was [Decision] made?”

### GEPA-Optimized Prompt – Task 1

Act as a clinical intake specialist. Your task is to extract the core medical concern from a patient’s narrative and rephrase it into a single, professional *Clinician Question*.

#### Follow these steps:

1. **Identify the Medical Catalyst:** Locate primary symptoms (e.g., “numbness,” “chest pain”), specific events (e.g., “rat bite,” “surgery”), or medications mentioned.
2. **Filter the “Noise”:** Remove all emotional context, pleasantries, expressions of gratitude, and conversational fillers.
3. **Synthesise the Query:** Formulate the question focused on the clinical objective—seeking a diagnosis (*What caused X?*), causality (*Was X caused by Y?*), or triage (*Should I take action?*).
4. **Maintain Terminology:** Use the patient’s specific language for symptoms and medications but organise it with professional structure.



### 5. Standardise the Phrasing:

- Use simple starters: “What caused...”, “Was the...”, “Should I...”, “Is it expected...”, “Why was...”.
- Convert vague worries like “What happened?” into “What did the findings show?” or “What caused these symptoms?”.
- Convert “Is that normal?” to “Is it expected for [symptom] to occur?”.
- Ensure the final output is a single, concise sentence a clinician can address immediately.

procedural details without outcome relevance, or redundant information.

### Strategy by Question Type:

- *Recovery / “What do I do?”*: Include prognosis, follow-up plan, and red flags.
- *“Is it connected?”*: Include positive evidence and significant negatives used to rule out connections.
- *“Why this procedure?”*: Focus on failure of prior treatments, worsening lab trends, and obstructions found.

**Output:** Provide only a list of the essential sentence IDs.

## Task 2 – Evidence Identification

### Base Prompt – Task 2

Given a patient’s medical question and a numbered set of sentences from a clinical note excerpt, identify the sentences that are essential to answering the question.

A sentence is *essential* if removing it would leave the answer to the patient’s question incomplete — it directly provides clinical evidence (findings, procedures, results, reasoning, or recommendations) that addresses what the patient asked.

The goal is precision: include every sentence without which the answer would be incomplete, but exclude sentences that merely provide context or are unrelated.

### GEPA-Optimized Prompt – Task 2

You are a reliable expert clinician. Given a patient’s medical narrative and a numbered set of sentences from a clinical note excerpt, identify all sentence IDs that are essential to answering the question.

**Definition of “Essential”:** A sentence is essential if its removal results in an incomplete answer. It must provide direct clinical evidence, including:

1. **Findings & Results:** Specific test results, diagnoses, or negative findings that rule out or confirm a condition.
2. **Clinical Reasoning:** The rationale behind a diagnosis or treatment path, including evidence of clinical deterioration that justified a procedure.
3. **Recommendations & Action Plans:** Specific follow-up instructions, including doctor names or specialties for further care.
4. **Safety Information (Red Flags):** Warning signs the patient is told to monitor and report immediately.

**Exclusion Criteria:** Exclude sentences providing only general context, hospital pleasantries, technical

## Task 3 – Answer Generation

### Base Prompt – Task 3

You are a medical expert tasked with answering a patient’s question using only information explicitly stated in the provided clinical note excerpt.

**Task:** Using only the essential sentences identified by the provided indices, write a response that directly and faithfully addresses the patient’s question.

### Rules:

- Use only facts explicitly stated in the clinical note. Do not speculate or infer beyond what is documented.
- If the note does not fully answer the question, answer what can be supported and acknowledge the gap. Do not fabricate information.
- Write in clear, simple language a patient or family member can understand.
- Do not include citations, sentence numbers, or references.
- Write a 3rd-person, concise, plain prose response directly addressing the patient’s question.
- Keep the answer under 75 words (~5 sentences).

### GEPA-Optimized Prompt – Task 3

You are a medical expert assistant. Extract and translate key clinical information from a patient’s medical record into a concise, medically precise explanation for the patient or their family.

### Task Instructions:

1. **Strict Evidence Filtering:** Use *only* the sentences corresponding to the provided `evidence_ids`.
2. **Focus on Clinical Reasoning:** Explain *why* medical actions were taken by connecting phys-

ical exam findings and diagnostic results to treatments.

3. **Symptom Documentation:** Incorporate all symptoms and physical exam findings mentioned in the filtered evidence.
4. **Synthesise Findings:** Do not include specific lab values or exact dosages; use interpretations instead (e.g., “elevated muscle enzymes” rather than “CK peaking at 1405”).
5. **Closing the Loop:** Directly address the patient’s concern with an evidence-based answer.

#### Strict Rules:

- **Factuality:** Do not use outside medical knowledge.
- **Tone & Perspective:** Write in the 3rd person. Use simple, clear sentences while retaining professional terminology.
- **Format:** Max 75 words; no citations or sentence numbers; single seamless paragraph.
- **Goal of Care:** Explicitly mention any transition to comfort measures or major shift in the direction of care if evidenced.

## Task 4 – Evidence Alignment

### Base Prompt – Task 4

Ground each sentence of a clinical answer to the specific sentence(s) in the clinical note excerpt that directly support it.

#### Alignment Rules:

- Operate at the answer-sentence level: each answer sentence maps to zero, one, or more note sentences.
- Only link a note sentence when it provides direct, specific support — not merely topical overlap.
- Alignments are many-to-many: one note sentence may support multiple answer sentences and vice versa.
- Answer sentences derived entirely from clinical knowledge outside the excerpt map to an empty list.
- Cite only note sentences that genuinely ground the claim in the answer sentence.

### GEPA-Optimized Prompt – Task 4

You are a clinical data analyst performing a high-fidelity audit of patient-facing explanations. Map simplified *Answer Sentences* to original *Note Sentences* using the standard of **Minimal Sufficient Proof**.

#### Core Principles:

1. **Minimal Sufficient Proof:** Select the small-

est set of note sentences that together provide 100% factual support. Apply the *Redundancy Test*: if removing a note sentence leaves the claim fully proven, remove it.

2. **Strict Claim Decomposition:** Break each answer sentence into atomic claims. Every claim must have an explicit corresponding fact in the note; if any claim is missing or only implied, the alignment for that sentence is [].
3. **Cross-Sectional Mapping:** If an answer sentence uses both technical and layman terms, select both the technical evidence and the summary evidence — but do not include underlying data when a single discharge instruction sentence already covers the claim.

#### Domain-Specific Rules:

- **Cardiac rule-outs:** Require both diagnostic data (troponins/EKGs) and the physician’s conclusion.
- **Musculoskeletal pain:** Require physical exam findings and the clinical conclusion.
- **Follow-up instructions:** Must identify doctor/specialty, timeframe/test, and purpose; incomplete matches align to [].
- **Headers:** Include only if the header is the sole source of the specific diagnosis in the answer sentence.

**Output Format:** Precede alignments with a **Reasoning** section listing atomic claims and justifying the minimal evidence set, followed by a JSON-style dictionary: {"1": [i, j], "2": [], "3": [k]}.

## A.2. LLM Judge Prompts for GEPA Optimization

For generative tasks (Tasks 1 and 3), GEPA uses an LLM judge to generate concise textual feedback by comparing the student’s output against the gold reference, guiding the teacher model in proposing improved prompt candidates.

#### Judge Prompt — Task 1: Q. Interpretation

You are an expert Medical QA evaluator. You are evaluating a system's ability to transform a free-text, patient-authored narrative into a clear and concise clinician-interpreted question. The goal is to maximize the similarity between the predicted clinical question and the gold question.

*Feedback guidelines:*

- Should be general and not focused on this sample.
- Should be concise and focused only on why the prediction differs from the gold.
- Should guide the system on how to improve its performance on this task in general.
- Do not include samples from the provided narrative.
- Keep it under 20 words.

#### Judge Prompt — Task 3: Answer Generation

You are an expert Medical QA evaluator. You are evaluating a system's ability to answer a patient-authored question using only information explicitly stated in the provided clinical note excerpt. The goal is to maximize the similarity between the predicted answer and the gold answer.

*Feedback guidelines:*

- Should be general and not focused on this sample.
- Should be concise and focused only on why the prediction differs from the gold.
- Should guide the system on how to improve its performance on this task in general.
- Do not include samples from the provided narrative.
- Keep it under 20 words.

# TAMU-NLP at ArchEHR-QA 2026: Grounded Clinical QA with Evidence Identification and Intent-Aware Answer Generation

Xinqi Su<sup>1</sup>, Rongrong Wang<sup>2</sup>, Sunyang Fu<sup>2</sup>, Hongfang Liu<sup>2</sup>, Ruihong Huang<sup>1</sup>

<sup>1</sup>Texas A&M University, College Station, TX, USA

<sup>2</sup>UTHealth Houston, Houston, TX, USA

{suxinqi666, huangrh}@tamu.edu

{Rongrong.Wang, Sunyang.Fu, Hongfang.Liu}@uth.tmc.edu

## Abstract

Electronic Health Records (EHRs) contain rich clinical information and provide an important data source for medical question answering. However, generating reliable answers grounded in patient-specific clinical evidence remains challenging. In this work, we participate in the ArchEHR-QA 2026 shared task and focus on Subtask 2 (Evidence Identification) and Subtask 3 (Answer Generation). For evidence identification, we explore both traditional learning-to-rank methods and large language models (LLMs), and propose a two-stage LLM framework that improves prediction stability through few-shot prompting and self-reflection reasoning. For answer generation, we design an intent-aware few-shot prompting framework to generate concise answers grounded in clinical evidence. Experimental results show that our approach achieves strong performance despite limited training data. On the official leaderboard, our system ranks 5th in Subtask 2 and 2nd in Subtask 3. These results demonstrate that combining evidence-driven reasoning with the generative capabilities of LLMs is an effective approach for EHR-based clinical question answering.

**Keywords:** Electronic Health Records (EHR), Clinical Question Answering, Evidence Identification, Answer Generation

## 1. Introduction

Electronic Health Records (EHRs) digitally store and manage patient health information, including medical history, laboratory results, medications, and clinical procedures, providing essential data for clinical decision-making and research. With the widespread adoption of EHR systems and patient portals, clinicians receive increasing numbers of patient inquiries, many of which arise from patients reviewing and questioning their health information (Cronin et al., 2015; Irizarry et al., 2015). These inquiries not only increase clinicians' workload but also consume substantial clinical resources.

Although prior work has explored automated health question answering (QA), such as clinical and biomedical QA systems (Ben Abacha and Demner-Fushman, 2019; Pampari et al., 2018), relatively limited attention has been paid to leveraging patient-specific clinical records and grounding answers in explicit clinical evidence. Generating personalized, evidence-supported responses from EHR data therefore represents a promising direction for improving clinical efficiency and providing patients with reliable health information.

The ArchEHR-QA 2026 shared task evaluates EHR-based question answering systems through four subtasks emphasizing evidence grounding and clinical interpretability. Subtask 1 (Question Interpretation) converts patient free-text questions into concise clinical queries understandable to clinicians. Subtask 2 (Evidence Identification) extracts key evidence sentences from clinical notes required to answer the question. Subtask 3 (Answer Gen-

eration) produces concise answers grounded in the selected evidence. Subtask 4 (Evidence Alignment) evaluates the alignment between answers and their supporting evidence.

In this work, we focus on Subtasks 2 and 3, which form a natural reasoning pipeline in EHR question answering: evidence identification determines the information basis for answer generation, while answer generation produces patient-oriented explanations grounded in the selected evidence (Ben Abacha and Demner-Fushman, 2019).

For Subtask 2, we explore both traditional learning-to-rank methods (Cao et al., 2007; Burges, 2010) and large language models (LLMs) for evidence identification. We propose a two-stage LLM-based evidence identification framework, which performs initial predictions using few-shot prompting and improves prediction stability through self-reflection reasoning and ensemble voting.

For Subtask 3, we adopt a few-shot intent-aware prompting framework based on LLMs to generate concise and evidence-grounded answers. By leveraging the reasoning capability of LLMs, the system can produce answers consistent with clinical evidence even under limited training data.

Our approach achieves competitive performance. On the official ArchEHR-QA 2026 leaderboard, our system ranks 5th in Subtask 2 (Evidence Identification) and 2nd in Subtask 3 (Answer Generation). These results demonstrate that combining evidence-driven reasoning strategies with the generative capabilities of large language models provides an effective solution for EHR-based clinical question answering.

## 2. Related Work

### 2.1. Evidence Identification

Evidence identification aims to extract a minimal set of sentences from clinical notes that support answer generation. Early approaches relied on rule-based systems or keyword matching, such as evidence extraction using regular expressions or ontology-based medical terms (Demner-Fushman and Lin, 2007). With the development of deep learning, supervised sentence-level models were introduced, treating evidence identification as a binary classification task to determine whether a candidate sentence supports the answer (Chakraborty et al., 2020). These models often extract entities and relations as features by performing information extraction tasks including clinical named entity recognition (NER) and relation extraction (RE). Recently, large language models (LLMs) have been applied to evidence identification, enabling zero-shot or few-shot evidence prediction through strong semantic understanding (Singhal et al., 2023). However, most existing studies focus on biomedical literature or general medical QA, while evidence retrieval from patient-specific EHR data remains underexplored, particularly under small datasets where balancing precision and recall is challenging.

### 2.2. Answer Generation

The goal of answer generation is to produce concise and accurate responses grounded in evidence. Early medical QA systems mainly relied on information retrieval or template-based generation, where retrieved evidence snippets were directly assembled into answers, often lacking fluency and flexibility (Ben Abacha and Zweigenbaum, 2015; Ben Abacha and Demner-Fushman, 2019). With the development of deep learning, neural generation models have been increasingly applied, learning relationships among questions, evidence, and answers to produce more natural responses (Pampari et al., 2018; Jin et al., 2019; Lee et al., 2020). More recently, large language models (LLMs) have demonstrated strong capabilities, generating coherent responses under zero-shot or few-shot settings (Singhal et al., 2023; Nori et al., 2023). However, most existing studies focus on medical exam questions or biomedical literature, while answer generation for patient-specific EHR data remains underexplored, particularly under small-scale training settings where maintaining both answer reliability and evidence traceability is challenging.

## 3. Data and Evaluation Metrics

### 3.1. Data

The ArchEHR-QA 2026 dataset (Soni and Demner-Fushman, 2026b,a) is organized at the case level.

The development set contains 20 cases (Cases 1–20). Each case consists of the following components:

- **Patient Narrative:** a free-text description of the patient’s symptoms or health concerns from the patient perspective.
- **Clinician Question:** a concise clinical question formulated by a clinician based on the patient narrative.
- **Patient–Question Phrases:** key phrases in the narrative annotated as relevant to the patient’s question.
- **Clinical Note Excerpts:** sentence-level excerpts from clinical notes, where each sentence is labeled as *essential*, *supplementary*, or *not relevant*. The notes are derived from the MIMIC database.
- **Clinician Answer:** a reference answer written by a clinician. Each answer sentence is supported by zero or more sentences from the clinical note excerpts.
- **Case Metadata:** metadata including the case ID, clinical specialty, and EHR source.

In addition to the development set, the dataset includes Cases 21–120, which lack annotated evidence labels for the clinical note excerpts. To address the lack of annotated training data, we employ a large language model (LLM) to automatically annotate evidence sentences from the clinical note excerpts based on the clinician answers. The resulting LLM-annotated data are used as training data for Task 2 (Evidence Identification), while the original development set (Cases 1–20) is used as the validation set. Furthermore, the official test set for Task 2 and Task 3 includes 47 additional cases (Cases 121–167), which follow a structure similar to the development set but do not include annotated evidence labels or clinician answers, and are used solely for evaluation.

Although the original annotations include three relevance labels, we cast the task as a binary classification problem since Task 2 evaluation and Task 3 clinician answers primarily rely on *essential* evidence. Specifically, only *essential* sentences are treated as gold evidence, while the other two categories are merged into *not-essential*. In the development set, the label distribution consists of 121 *essential* and 307 *not-essential* sentences, reflecting a class imbalance.

### 3.2. Evaluation Metrics

**Task 3: Answer Generation.** We evaluate answer generation using a diverse set of automatic



metrics that capture different aspects of answer quality. **BLEU** and **ROUGE** measure n-gram overlap between generated and reference answers. **BERTScore** evaluates semantic similarity between generated and reference texts. **SARI** (Xu et al., 2016) measures text editing quality by comparing added, deleted, and retained n-grams with respect to the reference. **AlignScore** (Rogers et al., 2023) assesses semantic and factual alignment using a trained scoring model that evaluates the consistency between generated and reference answers. **MEDCON** (Yim et al., 2023) evaluates medical consistency by checking whether clinically relevant information is preserved without contradiction. Together, these metrics provide a comprehensive evaluation of lexical similarity, semantic alignment, content faithfulness, and clinical correctness.

## 4. Methods

In this section, we describe in detail the approaches we explored for Task 2 and Task 3.

### 4.1. Task 2: Evidence Identification

#### 4.1.1. Task Formulation

Given a clinician question  $q$ , a patient narrative  $p$ , and a set of candidate sentences extracted from clinical notes  $S = \{s_1, s_2, \dots, s_n\}$ , the goal of evidence identification is to identify sentences that provide key evidence for answering the question.

Although the original annotations contain three labels (*essential*, *supplementary*, and *not-essential*), only *essential* sentences are treated as gold evidence during evaluation. We therefore formulate the task as a sentence-level relevance prediction problem:  $y_i \in \{0, 1\}$ , where  $y_i = 1$  indicates that sentence  $s_i$  is an essential evidence sentence.

Since each case contains multiple candidate sentences, we further model the task as a sentence ranking problem by learning a scoring function

$$\text{score}_i = f(q, p, s_i), \quad s_i \in S \quad (1)$$

such that evidence sentences receive higher scores than non-evidence sentences.

#### 4.1.2. Ranking-based Evidence Identification

**Evidence Ranking Model** To model the relevance between the question and candidate sentences, we adopt a cross-encoder architecture. Unlike bi-encoders that encode the query and sentence independently, the cross-encoder jointly encodes the question–sentence pair, enabling fine-grained token-level interactions (Reimers and Gurevych, 2019; Nogueira and Cho, 2019). Since each case contains only a small number of candidate sentences, the additional computational cost remains manageable.

We initially experimented with incorporating the clinical specialty and the patient narrative  $p$  into the query representation. However, empirical results show little improvement compared with using only the clinician question  $q$ . This is likely because the clinician question already summarizes the key information from the patient narrative, while the narrative often contains additional background details that are less relevant for evidence selection. Therefore, we use only the question  $q$  as the query input in the final model. For each question–sentence pair  $(q, s_i)$ , the input sequence is constructed as  $x_i = [\text{CLS}] \ q \ [\text{SEP}] \ s_i \ [\text{SEP}]$ . The sequence is fed into an encoder  $f_\theta$ , and the hidden representation of the  $[\text{CLS}]$  token is used as the sentence-pair representation:

$$h_i = f_\theta(x_i)_{[\text{CLS}]} \quad (2)$$

A linear layer is then applied to compute the relevance score:

$$\text{score}_i = W h_i + b \quad (3)$$

**Learning Objectives** We explore three learning strategies: **pointwise classification**, **pairwise ranking**, and **listwise ranking**, and further investigate **hybrid objectives** that combine classification and ranking signals.

**Pointwise Classification** We first model the task as a sentence-level binary classification problem. The probability of a sentence being evidence is computed as

$$p_i = \sigma(\text{score}_i). \quad (4)$$

The training objective is the binary cross-entropy (BCE) loss:

$$L_{\text{point}} = - \sum_i [y_i \log p_i + (1 - y_i) \log(1 - p_i)]. \quad (5)$$

Since evidence sentences are much fewer than non-evidence sentences, we apply weighted BCE loss to alleviate class imbalance (Lin et al., 2017).

**Pairwise Ranking** To learn the relative ordering of sentences, we adopt pairwise learning-to-rank. For each case, we construct positive–negative sentence pairs  $(s^+, s^-)$ . To reduce training difficulty, we perform hard negative mining, selecting the top- $k$  hardest negative sentences (i.e., those with the highest predicted scores). The margin ranking loss is defined as:

$$L_{\text{pair}} = \max(0, m - \text{score}_{\text{pos}} + \text{score}_{\text{neg}}). \quad (6)$$

This objective encourages

$$\text{score}_{\text{pos}} > \text{score}_{\text{neg}} + m. \quad (7)$$

thereby improving the model’s ability to distinguish evidence sentences from hard negatives.

**Listwise Ranking** We also explore listwise ranking, which learns sentence ranking at the case level. Given model scores  $\text{score}_i$ , we compute a probability distribution:

$$P_i = \frac{\exp(\text{score}_i)}{\sum_j \exp(\text{score}_j)}. \quad (8)$$

The target distribution assigns probability mass only to gold evidence sentences:

$$G_i = \frac{y_i}{\sum_j y_j}. \quad (9)$$

The listwise loss is defined as

$$L_{\text{list}} = - \sum_i G_i \log P_i. \quad (10)$$

**Hybrid Objectives** To combine the benefits of classification and ranking signals, we further explore hybrid objectives:

**Pointwise + Pairwise:**

$$L = \alpha L_{\text{point}} + (1 - \alpha) L_{\text{pair}}. \quad (11)$$

**Pointwise + Listwise:**

$$L = \alpha L_{\text{point}} + (1 - \alpha) L_{\text{list}}. \quad (12)$$

where  $\alpha$  is a weighting coefficient.

**Inference** During inference, the model computes a score  $\text{score}_i$  or probability  $p_i$  for each candidate sentence and applies several strategies to determine the number of evidence sentences.

**Fixed Threshold** We first select evidence sentences using a fixed probability threshold:

$$\hat{E} = \{s_i \mid p_i \geq \tau\}. \quad (13)$$

where the threshold  $\tau$  is tuned on the development set to maximize Strict Micro-F1. If no sentence exceeds the threshold, the sentence with the highest score is selected.

**Score-Gap Selection** To adaptively determine the number of evidence sentences, we adopt a score-gap strategy. Candidate sentences are first sorted by their scores:  $\text{score}_{(1)} \geq \text{score}_{(2)} \geq \dots \geq \text{score}_{(n)}$ . We then compute the score difference between adjacent sentences:

$$\text{gap}_i = \text{score}_{(i)} - \text{score}_{(i+1)}. \quad (14)$$

The largest gap typically indicates the boundary between evidence and non-evidence sentences:

$$k = \arg \max_i \text{gap}_i. \quad (15)$$

The top- $k$  sentences are selected as predicted evidence:  $\hat{E} = \{s_{(1)}, \dots, s_{(k)}\}$ .

**Adaptive Soft Dynamic- $k$**  We further propose an Adaptive Soft Dynamic- $k$  method that determines the number of evidence sentences based on prediction uncertainty. Sentence scores are first normalized with softmax:

$$p_i = \frac{\exp(\text{score}_i)}{\sum_j \exp(\text{score}_j)}. \quad (16)$$

We then compute the normalized entropy of the distribution:

$$H_{\text{norm}} = - \frac{\sum_i p_i \log p_i}{\log(n)}. \quad (17)$$

A dynamic threshold is defined as

$$\tau = \tau_0 + \lambda H_{\text{norm}}, \quad (18)$$

where  $\tau_0$  and  $\lambda$  are tuned on the development set. Sentences are sorted by probability and accumulated until

$$\sum_{i=1}^k p_i \geq \tau. \quad (19)$$

This yields a dynamic number of evidence sentences  $k$ . When the prediction distribution is concentrated (low entropy), only a few high-confidence sentences are selected; when it is more dispersed (high entropy), more sentences are retained. This helps preserve important evidence and provides more reliable support for answer generation.

#### 4.1.3. LLM-based Evidence Identification

To explore the capabilities of large language models (LLMs) for evidence identification, we design a two-stage LLM framework that combines retrieval-based few-shot prompting and self-reflection reasoning to improve prediction accuracy and stability.

**Retrieval-based Few-shot Prompting** To guide the model toward consistent annotation patterns, we adopt retrieval-based few-shot prompting. Instead of randomly selecting examples, we retrieve several annotated cases from the development set based on semantic similarity to the current query.

Each example contains the patient narrative, clinician question, clinical note, and labels. These examples provide explicit supervision signals that help the model determine whether a sentence constitutes key evidence.

**Stage 1: Initial Evidence Prediction** In the first stage, we use Gemini-2.5-Pro to perform initial evidence identification. The model receives the retrieved few-shot examples, patient narrative, clinician question, and clinical note, and predicts structured labels for each sentence:

$$Y^{(1)} = \{y_1^{(1)}, y_2^{(1)}, \dots, y_n^{(1)}\}. \quad (20)$$

Each sentence is labeled as *essential* or *not-essential*.

**Stage 2: Self-Reflection Refinement** Single-pass LLM predictions may miss important evidence or over-predict irrelevant sentences. To mitigate this issue, we introduce a self-reflection stage. The model receives the initial prediction  $Y^{(1)}$  and re-evaluates its decisions by checking whether key medical evidence was missed or incorrectly labeled. After reflection, the model produces refined predictions:

$$Y^{(2)} = \{y_1^{(2)}, y_2^{(2)}, \dots, y_n^{(2)}\}. \quad (21)$$

The final output is defined as  $Y = Y^{(2)}$ .

**Ensemble Voting** To further improve robustness, we perform ensemble voting over multiple LLM predictions. For each sentence, we count how many times it is labeled as *essential*. A sentence is predicted as essential if its vote count exceeds a predefined threshold  $T$ :

$$y_i = \begin{cases} 1 & v_i \geq T \\ 0 & \text{otherwise} \end{cases} \quad (22)$$

where  $v_i$  denotes the number of *essential* votes. Adjusting  $T$  allows the system to flexibly balance precision and recall in evidence identification.

## 4.2. Task 3: Answer Generation

Given a clinician question  $q$ , a patient narrative  $n$ , and the evidence  $E = \{e_1, e_2, \dots, e_k\}$  identified in Section 4.1, the goal of answer generation is to produce an answer  $a$  grounded in the evidence.

Clinical questions often reflect different information needs, such as treatment planning, prognosis assessment, or interpretation of abnormal laboratory results. To better support clinical reasoning and improve answer quality, we incorporate question intent information and retrieve intent-aligned demonstrations to guide generation. This intent-aware strategy provides reasoning patterns that match the question type, improving consistency between generated answers and clinical evidence. Our framework consists of three stages: (1) intent identification, (2) intent-aware example retrieval, and (3) evidence-grounded answer generation.

### 4.2.1. Intent Identification

We formulate intent prediction as a classification task. For each case  $x = (n, q)$ , we define a taxonomy of 12 clinical question intents (e.g., treatment plan, medication safety, prognosis, and laboratory abnormality explanation). Given a question, the model predicts a primary intent  $I$  and optionally a secondary intent  $S$ . The predicted intents are generated by the DeepSeek-V3 model, yielding a structured representation  $x = (n, q, I, S)$ .

### 4.2.2. Intent-Aware Example Retrieval

To support in-context learning, we construct an example pool  $P$  with intent annotations. For a test sample  $x_i$  with primary intent  $I_i$  and secondary intent  $S_i$ , we first retrieve candidate examples that share the same primary intent:

$$C_i = \{x_j \in P \mid I_j = I_i\}. \quad (23)$$

If the number of retrieved candidates is fewer than  $K$ , we further include examples that match the secondary intent  $S_i$ . The example pool is constructed from Cases 1–120, where each case is annotated with intent labels. The pool covers all predefined intent categories, ensuring that each intent type has sufficient example support.

We then rank candidates using semantic similarity. Each sample is represented as  $t = I \oplus q \oplus n$ , where  $\oplus$  denotes text concatenation. Let  $\mathbf{v}_q$  and  $\mathbf{v}_j$  denote the embeddings of the query and candidate example. Similarity is computed as

$$\text{sim}(q, x_j) = \frac{\mathbf{v}_q \cdot \mathbf{v}_j}{\|\mathbf{v}_q\| \|\mathbf{v}_j\|}. \quad (24)$$

The top- $K$  most similar examples are selected:

$$D_i = \text{TopK}_{x_j \in C_i} \text{sim}(x_i, x_j). \quad (25)$$

### 4.2.3. Evidence-Grounded Answer Generation

Finally, we generate answers using Gemini-2.5-Pro conditioned on the question, evidence sentences, intent information, and retrieved demonstrations. The generation function can be written as

$$a = g(n, q, E, D). \quad (26)$$

Each demonstration in  $D$  contains the patient narrative, clinician question, evidence sentences, and a reference answer, illustrating clinically grounded reasoning. Incorporating evidence and intent information encourages the model to generate answers consistent with the supporting evidence.

## 5. Results

In this section, we present the implementation details of our approaches and report experimental results on the development set as well as the official evaluation results.

### 5.1. Implementation Details

All experiments are implemented in PyTorch with the Transformers library and run on two NVIDIA RTX 3090 GPUs. For ranking-based evidence

| Method               | Strict        |               |               |               |               |               | Lenient       |               |               |               |               |               |
|----------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                      | Macro         |               |               | Micro         |               |               | Macro         |               |               | Micro         |               |               |
|                      | P             | R             | F1            | P             | R             | F1            | P             | R             | F1            | P             | R             | F1            |
| Pointwise            | 0.5113        | 0.3692        | 0.3886        | 0.5000        | 0.3471        | 0.4098        | 0.5472        | 0.3364        | 0.3826        | 0.5476        | 0.3129        | 0.3983        |
| Pairwise Ranking     | 0.4463        | <b>0.8566</b> | 0.5594        | 0.3820        | <b>0.8430</b> | 0.5258        | 0.5201        | <b>0.8225</b> | <b>0.6104</b> | 0.4532        | <b>0.8231</b> | 0.5845        |
| Listwise Ranking     | 0.4750        | 0.7205        | 0.5462        | 0.4750        | 0.6281        | 0.5409        | 0.5437        | 0.6920        | 0.5820        | 0.5437        | 0.5918        | 0.5668        |
| Pointwise + Pairwise | 0.5214        | 0.6758        | <b>0.5617</b> | 0.5214        | 0.6033        | <b>0.5594</b> | 0.6000        | 0.6634        | 0.6017        | 0.6000        | 0.5714        | <b>0.5854</b> |
| Pointwise + Listwise | <b>0.5500</b> | 0.6278        | 0.5554        | <b>0.5500</b> | 0.5455        | 0.5477        | <b>0.6250</b> | 0.5881        | 0.5781        | <b>0.6250</b> | 0.5102        | 0.5618        |

Table 1: Results on Task 2: Evidence Identification on the development set under Strict and Lenient evaluation settings. We report Macro and Micro Precision (P), Recall (R), and F1.

| Method                     | P             | R             | F1            |
|----------------------------|---------------|---------------|---------------|
| Fixed Threshold            | 0.5214        | 0.6033        | 0.5594        |
| Score-Gap Selection        | 0.4875        | <b>0.6446</b> | 0.5552        |
| Adaptive Soft Dynamic- $k$ | <b>0.5469</b> | 0.5800        | <b>0.5623</b> |

Table 2: Effect of inference strategies on Task 2 under the Strict setting (development set). We report Precision (P), Recall (R), and Micro-F1.

identification, we adopt a cross-encoder architecture based on BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext. We also experiment with several alternative backbones, including DeBERTa-v3-large, DeBERTa-v3-base, BioClinicalBERT, and BioBERT-base-cased-v1.2. However, these models yield similar performance on the development set, so we report the results using PubMedBERT.

The model is trained for 10 epochs using the AdamW optimizer with a learning rate of  $2 \times 10^{-5}$  and weight decay of 0.01. The maximum input length is set to 384, and gradient clipping with a norm of 1.0 is applied. For the hybrid loss, the weight  $\alpha$  is set to 0.1 and margin is set to 0.2. To ensure reproducibility, all experiments use a fixed random seed of 42. For LLM-based methods, we access the models through the official API.

## 5.2. Task 2: Evidence Identification

### 5.2.1. Effect of Ranking Objectives on Evidence Identification

Table 4.2.3 reports the performance of different evidence identification methods on the development set under both Strict and Lenient evaluation settings, including Macro and Micro Precision, Recall, and F1. All experiments use a fixed threshold (0.4) during inference for evidence selection. The results show that pointwise classification performs the worst overall. Although it achieves relatively high Micro Precision (0.5000) under the Strict setting, its Recall is much lower (0.3471), resulting in a Strict Micro-F1 of only 0.4098. This suggests that modeling evidence identification as independent sentence-level classification is insufficient to recover all key evidence sentences, since the task

inherently requires comparing multiple candidate sentences within the same case.

In contrast, ranking-based methods substantially improve recall. Pairwise ranking achieves the highest Strict Micro Recall (0.8430), while listwise ranking also outperforms pointwise classification. This indicates that modeling the relative relevance among candidate sentences within each case better matches the nature of the task.

Combining classification and ranking objectives further improves the precision-recall balance. For example, Pointwise + Pairwise increases Precision while maintaining high Recall, and Pointwise + Listwise achieves the best Strict Micro-F1 (0.5477). Similar trends are observed under the Lenient setting, where Pointwise + Pairwise obtains the highest Lenient Micro-F1 (0.5854).

Overall, these results suggest that modeling evidence identification as a ranking problem is more suitable, and integrating classification and ranking signals further improves performance, particularly under limited training data.

### 5.2.2. Inference Strategy Analysis

Table 4.2.3 compares three evidence selection strategies under the Strict evaluation setting: Fixed Threshold, Score-Gap Selection, and Adaptive Soft Dynamic- $k$ . All strategies are applied on top of the Pointwise + Pairwise model during inference. Although the overall F1 scores are similar, the methods exhibit different precision-recall trade-offs.

The Fixed Threshold method achieves relatively higher recall but lower precision, indicating that a fixed threshold tends to retain more candidate sentences, increasing recall while introducing more false positives. Score-Gap Selection achieves the highest recall (0.6446) but the lowest precision (0.4875). This is consistent with the intuition behind the method: automatically determining the number of evidence sentences based on score gaps tends to retain more potentially relevant sentences, which improves recall but also increases false positives.

In contrast, Adaptive Soft Dynamic- $k$  achieves the best Strict Micro-F1 and the highest precision while maintaining stable recall. This suggests that dynamically adjusting the number of evidence

| Method                              | Strict        |               |               |               |               |               | Lenient       |               |               |               |               |               |
|-------------------------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                                     | Macro         |               |               | Micro         |               |               | Macro         |               |               | Micro         |               |               |
|                                     | P             | R             | F1            | P             | R             | F1            | P             | R             | F1            | P             | R             | F1            |
| Few-shot                            | 0.8026        | 0.8269        | 0.7917        | 0.7692        | 0.8264        | 0.7968        | 0.8231        | 0.7060        | 0.7351        | 0.8000        | 0.7075        | 0.7509        |
| Few-shot + Self-reflection          | <b>0.8523</b> | <b>0.8427</b> | <b>0.8252</b> | 0.7923        | <b>0.8512</b> | <b>0.8207</b> | <b>0.8773</b> | <b>0.7229</b> | <b>0.7679</b> | <b>0.8385</b> | <b>0.7415</b> | <b>0.7870</b> |
| Few-shot + Self-reflection + Voting | 0.8492        | 0.8416        | 0.8229        | <b>0.7969</b> | 0.8430        | 0.8193        | 0.8702        | 0.7169        | 0.7610        | 0.8359        | 0.7279        | 0.7782        |

Table 3: Ablation study of the LLM-based evidence identification framework on Task 2 evaluated on the development set. We report Macro and Micro Precision (P), Recall (R), and F1.

| Method         | P             | R             | F1            |
|----------------|---------------|---------------|---------------|
| Qwen3-Max      | 0.6944        | 0.8264        | 0.7547        |
| Deepseek-V3    | 0.6809        | 0.7934        | 0.7328        |
| Deepseek-R1    | 0.7500        | 0.7934        | 0.7711        |
| GPT-4.1        | 0.7966        | 0.7769        | 0.7866        |
| GPT-5.1        | <b>0.8103</b> | 0.7769        | 0.7932        |
| Gemini-2.5-Pro | 0.7923        | <b>0.8512</b> | <b>0.8207</b> |

Table 4: Effect of different LLMs on Task 2 on the development set under the Strict evaluation setting. We report Precision (P), Recall (R), and Micro F1.

sentences based on prediction uncertainty better adapts to varying evidence distributions across cases, resulting in a more balanced performance.

Overall, these results indicate that the evidence selection strategy during inference has a substantial impact on performance, and uncertainty-based dynamic selection provides a better balance between precision and recall.

### 5.2.3. LLM-based Evidence Identification

Table 4.2.3 presents results of the LLM-based evidence identification, including Few-shot prompting (with  $K = 5$  retrieved in-context examples), Few-shot + Self-reflection, and Few-shot + Self-reflection + Voting. Few-shot prompting alone achieves strong performance. Under the Strict setting, it obtains a Micro-F1 of 0.7968, substantially outperforming traditional ranking models. This suggests that structured in-context examples effectively guide LLMs to identify key evidence sentences.

Adding self-reflection further improves performance. Under the Strict setting, Micro-F1 increases from 0.7968 to 0.8207, with gains in both precision and recall. Similar improvements are observed under the Lenient setting. This indicates that the reflection stage allows the model to revisit its initial predictions and correct missing or redundant evidence. In contrast, adding voting after self-reflection does not yield further improvements. Although the results remain competitive, they are slightly lower than those obtained with self-reflection alone under both settings. A possible explanation is that predictions across runs tend to be highly similar, and majority voting may suppress

some infrequent but important evidence sentences, slightly reducing recall.

Overall, these results indicate that the main performance gains of LLM-based evidence identification come from self-reflection reasoning, rather than simple multi-run voting.

### 5.2.4. Comparison of LLM Backbones

Table 4.2.3 compares the performance of different LLM backbones on the evidence identification task. Gemini 2.5 Pro achieves the best results (Strict Micro-F1 = 0.8207) and the highest recall (0.8512), indicating stronger coverage in identifying clinically relevant evidence sentences. Given the imbalanced label distribution in the development set (121 *essential* vs. 307 *not-essential*), recall becomes particularly important for capturing sparse gold evidence. The higher recall of Gemini suggests that it is more effective at retrieving relevant evidence under this setting. The GPT models (GPT-4.1 and GPT-5.1) also perform well, with higher precision but slightly lower recall than Gemini, suggesting a more conservative evidence selection strategy that prioritizes precision over coverage. Among open-source models, DeepSeek-R1 performs best (Micro-F1 = 0.7711), while Qwen3-Max and DeepSeek-V3 perform slightly worse. This indicates that reasoning and instruction-following capabilities are particularly important for sentence-level evidence identification. Overall, stronger language understanding and reasoning abilities help models better distinguish key evidence from background information, especially under imbalanced label distributions.

## 5.3. Test Set Results

Since Few-shot + Self-reflection achieved the best performance on the development set, we adopt it as the final method for Subtask 2. Table 4.2.3 reports the official test results, where our system obtains an overall score of 61.4, ranking among the top systems in the shared task. Compared with higher-ranked systems such as Neural and OptiMed, our method maintains relatively high Strict Micro Recall (70.0) but slightly lower precision. This indicates



| Method              | Overall     | Strict Micro |             |             | Lenient Micro |             |             | Strict Macro |             |             | Lenient Macro |             |             |
|---------------------|-------------|--------------|-------------|-------------|---------------|-------------|-------------|--------------|-------------|-------------|---------------|-------------|-------------|
|                     |             | P            | R           | F1          | P             | R           | F1          | P            | R           | F1          | P             | R           | F1          |
| Neural              | <b>63.7</b> | <b>60.2</b>  | 67.6        | <b>63.7</b> | <b>77.8</b>   | 67.6        | <b>72.4</b> | <b>64.3</b>  | 69.7        | <b>64.8</b> | <b>81.9</b>   | 69.7        | <b>73.1</b> |
| OptiMed             | 63.2        | 56.7         | 71.3        | 63.2        | 73.1          | 71.3        | 72.2        | 61.5         | 73.0        | 64.1        | 78.1          | 73.0        | 72.9        |
| UIC-AIHealth4All    | 62.9        | 59.3         | 67.0        | 62.9        | 74.3          | 67.0        | 70.4        | 61.8         | 69.2        | 63.0        | 77.8          | 69.2        | 70.5        |
| Yale-DM-Lab         | 61.9        | 52.3         | <b>75.7</b> | 61.9        | 68.0          | <b>75.7</b> | 71.6        | 56.7         | <b>75.5</b> | 61.1        | 73.8          | <b>75.5</b> | 70.4        |
| TAMU-NLP-Lab (Ours) | 61.4        | 54.6         | 70.0        | 61.4        | 74.4          | 70.0        | 72.2        | 61.4         | 70.3        | 60.6        | 78.9          | 70.3        | 70.7        |

Table 5: Official leaderboard Results on Task 2 on the test set. We report Strict and Lenient Micro and Macro Precision (P), Recall (R), and F1 scores.

| Method                           | Overall      | BLEU        | ROUGE-Lsum   | SARI         | BERTScore    | AlignScore   | MEDCON       |
|----------------------------------|--------------|-------------|--------------|--------------|--------------|--------------|--------------|
| LLM + Evidence + Few-shot        | 34.80        | 8.97        | 27.35        | 58.87        | 44.80        | 25.92        | <b>42.89</b> |
| LLM + Evidence + Few-shot Intent | <b>36.20</b> | <b>9.70</b> | <b>27.40</b> | <b>59.60</b> | <b>46.20</b> | <b>34.50</b> | 39.90        |

Table 6: Results on Task 3: Answer Generation on the test set. We report Overall score, BLEU, ROUGE-Lsum, SARI, BERTScore, AlignScore, and MEDCON.

| Method                                | Overall      | BLEU        | ROUGE-Lsum   | SARI         | BERTScore    | AlignScore   | MEDCON       |
|---------------------------------------|--------------|-------------|--------------|--------------|--------------|--------------|--------------|
| LLM + Evidence + Intent (Low-Recall)  | 35.74        | 8.54        | <b>27.81</b> | 59.48        | <b>46.77</b> | 31.85        | <b>39.98</b> |
| LLM + Evidence + Intent (High-Recall) | <b>36.20</b> | <b>9.70</b> | 27.40        | <b>59.60</b> | 46.20        | <b>34.50</b> | 39.90        |

Table 7: Results on Task 3: Answer Generation on the test set. We report Overall score, BLEU, ROUGE-Lsum, SARI, BERTScore, AlignScore, and MEDCON.

| Team                 | Overall     | BLEU       | ROUGE-Lsum  | SARI        | BERTScore   | AlignScore  | MEDCON      |
|----------------------|-------------|------------|-------------|-------------|-------------|-------------|-------------|
| WisPerMed            | <b>36.3</b> | <b>9.9</b> | <b>27.8</b> | 58.6        | <b>46.8</b> | 31.7        | 43.1        |
| TAMU-NLP-Lab (Ours)  | 36.2        | 9.7        | 27.4        | 59.6        | 46.2        | <b>34.5</b> | 39.9        |
| BIT.UA-AAUBS         | 35.6        | 8.6        | 26.4        | <b>60.0</b> | 45.0        | 30.2        | <b>43.2</b> |
| Neural               | 35.2        | 9.4        | 25.6        | 57.7        | 43.5        | 34.3        | 40.9        |
| HealthNLP_Retrievers | 34.6        | 7.0        | 25.4        | 59.2        | 43.8        | 33.6        | 38.7        |

Table 8: Official leaderboard results on ArchEHR-QA 2026 task 3 (Test Set).

that the system effectively covers key evidence sentences, although it may occasionally include some non-evidence sentences. Under the Lenient setting, our system achieves a Micro-F1 of 72.2 and a Lenient Macro-F1 of 70.7, suggesting stable performance across cases. Overall, the results show that our approach generalizes well to the shared task test set. Further improvements may come from enhancing precision in evidence selection.

## 5.4. Task 3: Answer Generation

### 5.4.1. Impact of Intent-Aware Few-Shot Retrieval

Table 4.2.3 compares two evidence-grounded answer generation settings: LLM + Evidence + Few-shot (with  $K = 3$  retrieved in-context examples) and LLM + Evidence + Few-shot Intent. Both methods use evidence sentences and few-shot demonstrations, while the latter additionally retrieves demonstrations based on question intent. Introducing intent information improves the Overall score from 34.80 to 36.20, indicating that the selection of few-shot demonstrations plays an important role in clinical answer generation. Compared with generic examples, intent-matched demonstrations provide

more effective guidance for generating high-quality responses. Across metrics, BLEU (8.97  $\rightarrow$  9.70) and ROUGE-Lsum (27.35  $\rightarrow$  27.40) show slight improvements, while BERTScore (44.80  $\rightarrow$  46.20) and AlignScore (25.92  $\rightarrow$  34.50) increase more substantially, suggesting better semantic alignment between generated and reference answers. In addition, SARI improves from 58.87 to 59.60, indicating better evidence integration and text restructuring. However, MEDCON slightly decreases (42.89  $\rightarrow$  39.90), suggesting that improved semantic relevance may come at the cost of stricter medical terminology consistency. Overall, intent-aware retrieval of few-shot demonstrations significantly improves semantic alignment and overall answer quality in clinical answer generation.

### 5.4.2. Impact of Evidence Recall on Answer Generation

Table 4.2.3 compares the Task 3 answer generation results under two evidence input settings: Low-Recall Setting and High-Recall Setting. These two settings are constructed by controlling the precision-recall trade-off in the evidence identification stage. Specifically, the Low-Recall setting applies

a stricter evidence selection criterion to retain only high-confidence sentences, leading to higher precision but lower recall. In contrast, the High-Recall setting adopts a more permissive selection strategy, including more candidate sentences, which increases recall but reduces precision.

Both methods use the same LLM + Evidence + Intent framework and differ only in the coverage of the input evidence. Overall, the High-Recall Setting achieves a higher Overall score (36.20 vs. 35.74), indicating that broader evidence coverage generally improves the quality of generated clinical answers. Across individual metrics, this setting performs better on BLEU, SARI, and AlignScore, with the largest improvement observed in AlignScore (31.85 → 34.50). This suggests that, for LLMs, incorporating more evidence can significantly enhance the semantic alignment between generated answers and reference answers. In contrast, the Low-Recall Setting performs slightly better on ROUGE-Lsum, BERTScore, and MEDCON, suggesting that a more selective evidence set may sometimes produce answers that are more focused and contain more compact medical concepts.

#### 5.4.3. Leaderboard Results

Table 4.2.3 presents the official leaderboard results on the ArchEHR-QA 2026 Subtask 3 test set. Our system (TAMU-NLP-Lab) ranks second with an Overall score of 36.2, only slightly below WisPerMed (36.3), demonstrating strong competitiveness. Notably, our system achieves the highest AlignScore (34.5) among all teams, indicating that the generated answers are most semantically aligned with the reference answers. This suggests that combining intent-aware generation with evidence grounding helps the model better capture the core information needs of clinical questions. Our system also maintains competitive performance on BLEU (9.7) and ROUGE-Lsum (27.4), showing strong surface-level similarity with the reference answers. However, our system does not achieve the best results on SARI and MEDCON. For example, BIT.UA-AAUBS obtains the highest SARI (60.0), and both BIT.UA-AAUBS and WisPerMed slightly outperform our system on MEDCON. This suggests that while our approach focuses on addressing question intent, there is still room for improvement in text editing quality and strict medical concept consistency.

## 6. Conclusion

This paper presents an evidence-grounded framework for clinical question answering and conducts a comprehensive evaluation on the ArchEHR-QA dataset. The task is decomposed into two stages:

evidence identification (Task 2) and answer generation (Task 3). For Task 2, we explore both ranking-based models and LLM-based evidence annotation. For Task 3, we introduce an intent-aware answer generation framework that retrieves few-shot demonstrations based on question intent to guide evidence-consistent responses. Experimental results show that incorporating LLM-based self-reflection significantly improves evidence identification, especially under limited training data, outperforming traditional ranking models. For answer generation, we find that evidence coverage and intent-aware few-shot retrieval are critical: higher evidence recall improves answer quality, while intent-based example retrieval enhances semantic alignment between generated and reference answers. In the ArchEHR-QA 2026 Shared Task, our system ranks 5th in Subtask 2 and 2nd in Subtask 3, achieving competitive performance across multiple metrics. Future work will explore tighter integration between evidence identification and answer generation to further improve end-to-end clinical QA performance.

## 7. Bibliographical References

- Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511.
- Asma Ben Abacha and Pierre Zweigenbaum. 2015. Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information Processing & Management*, 51(5):570–594.
- Christopher Burges. 2010. From ranknet to lambdamart to lambdamart: An overview. *Learning*, 11:81.
- Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of ICML*, pages 129–136.
- Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. 2020. Biomedbert: A pre-trained biomedical language model for qa and ir. In *Proceedings of COLING*, pages 669–679.
- Robert M Cronin, SE Davis, JA Shenson, Q Chen, ST Rosenbloom, and GP Jackson. 2015. Growth of secure messaging through a patient portal as a form of outpatient interaction across clinical specialties. *Applied Clinical Informatics*, 6(02):288–304.

- Dina Demner-Fushman and Jimmy Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103.
- Taya Irizarry, Annette DeVito Dabbs, and Christine Curran. 2015. Patient portals and patient engagement: a state of the science review. *Journal of Medical Internet Research*, 17(6):e148.
- 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.
- 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.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of ICCV*, pages 2980–2988.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.
- 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*.
- 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.
- Nils Reimers and Iryna Gurevych. 2019. Sentencebert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992.
- Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. 2023. Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, Sara Mahdavi, Jason Wei, Hyung Won Chung, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Sarvesh Soni and Dina Demner-Fushman. 2026a. [A dataset for addressing patient’s information needs related to clinical course of hospitalization](#). *Scientific Data*.
- 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.
- 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.
- Wen-wai Yim, Yajuan 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.

## A. Prompt for Evidence Identification

### A.1. Evidence Identification Prompt (Stage 1)

We use large language model to identify sentences that are essential for answering the clinician question. The model performs binary classification for each sentence: *essential* or *not-essential*. The prompt template is shown below.

Task:

For each sentence, decide whether it is ESSENTIAL to answer the clinician question.

Definitions:

- essential: directly necessary to answer the question.
- not-essential: not directly required.

Precision rule:

Only label as "essential" if removing the sentence would make answering incomplete or incorrect.

Rules:

- Think step-by-step internally but DO NOT output reasoning.
- Output JSON only.
- Include ALL sentence ids.
- Allowed labels: "essential" or "not-essential"

===== FEW-SHOT EXAMPLES =====

Patient Narrative:

{PATIENT\_NARRATIVE\_EXAMPLE}

Clinician Question:

{CLINICIAN\_QUESTION\_EXAMPLE}

Sentences:

1: {SENTENCE\_1}

2: {SENTENCE\_2}

3: {SENTENCE\_3}

...

Correct Output:

```
{
 "1": "essential",
 "2": "not-essential",
 "3": "essential"
}
```

===== END OF FEW-SHOT =====

Now classify:

Patient Narrative:

{PATIENT\_NARRATIVE}

Clinician Question:

{CLINICIAN\_QUESTION}

Sentences:

1: {SENTENCE\_1}

2: {SENTENCE\_2}  
3: {SENTENCE\_3}  
...

Output JSON only.

## A.2. Self-Reflection Prompt (Stage 2)

After the initial prediction, we perform a second-pass self-reflection step where the model reviews its previous decisions and corrects potential mistakes.

You previously classified the sentences as:

{INITIAL\_PREDICTION\_JSON}

Now carefully review your previous classification.

Check:

1. Is any truly essential sentence missing?
2. Is any labeled "essential" actually background?
3. Are essential sentences strictly necessary?

Revise if needed.

Rules:

- Think step-by-step internally but DO NOT output reasoning.
- Output JSON only.
- Include ALL sentence ids.
- Allowed labels: "essential" or "not-essential"

Patient Narrative:

{PATIENT\_NARRATIVE}

Clinician Question:

{CLINICIAN\_QUESTION}

Sentences:

1: {SENTENCE\_1}  
2: {SENTENCE\_2}  
3: {SENTENCE\_3}  
...

Output JSON only.

## B. Prompt for Answer Generation

### B.1. Prompt for Clinical Question Intent Classification

We use large language model to classify the clinical intent of each question. The exact prompt used for intent classification is shown below.

You must classify the primary intent of this clinical QA case.

Choose one primary intent from:

1. procedure\_justification
2. icu\_or\_monitoring\_necessity
3. treatment\_plan
4. medication\_safety
5. pathophysiology\_explanation



6. malignancy\_concern
7. lab\_abnormal\_concern
8. recovery\_duration
9. prognosis\_severity
10. activity\_clearance
11. persistent\_unexplained\_symptoms
12. neurologic\_mental\_status\_change

Also optionally choose a secondary intent (or null).

Also assign a risk level:  
low / moderate / high

Definitions:

- activity\_clearance = travel, flying, oxygen removal, return to work
- malignancy\_concern = fear of cancer
- pathophysiology\_explanation = asking why X caused Y
- lab\_abnormal\_concern = abnormal lab value severity
- recovery\_duration = asking how long recovery takes
- prognosis\_severity = asking how serious / survival outlook
- medication\_safety = tapering, toxicity, side effects
- persistent\_unexplained\_symptoms = chronic unexplained symptoms
- neurologic\_mental\_status\_change = confusion, dementia, seizures
- treatment\_plan = what should be done next
- procedure\_justification = why surgery/procedure done
- icu\_or\_monitoring\_necessity = why ICU or monitoring required

Patient Narrative:  
{PATIENT\_NARRATIVE}

Clinician Question:  
{CLINICIAN\_QUESTION}

Output JSON format:

```
{
 "primary": "...",
 "secondary": "...",
 "risk_level": "..."
}
```

## B.2. Answer Generation Prompt

You are a senior clinical specialist.

Your task is to generate a precise, evidence-grounded answer to the Clinician Question using ONLY the provided Clinical Note.

PRIMARY OPTIMIZATION GOALS:

A) Maximize lexical overlap with the Clinical Note:

Reuse the exact medical terms and phrasing from the note whenever possible. Prefer copying short clauses rather than paraphrasing.

B) Maximize clinically meaningful concept coverage:

Explicitly mention all key clinical concepts present in the note, including: diagnoses, symptoms, procedures, imaging or laboratory tests, medications, complications, anatomy, and clinically relevant numbers or timelines.

STRICT RULES:

- Use ONLY information documented in the Clinical Note.
- Do NOT add external knowledge.
- Do NOT speculate beyond the note.
- Prioritize essential sentences.
- Do NOT include sentence numbers, brackets, or citations.
- Use explicit clinical nouns rather than vague pronouns.
- Write in concise professional clinical register.
- Limit to 75 words maximum.
- Use 2-4 sentences.
- Avoid filler phrases.

COMPRESSION STRATEGY:

Combine related clinical facts into high-information sentences.  
Retain specific terminology such as procedure names, medication names,  
and diagnoses.  
Preserve original phrasing where possible to increase textual alignment.

===== FEW-SHOT EXAMPLES =====

Patient Question:  
{PATIENT\_QUESTION\_EXAMPLE}

Clinician Question:  
{CLINICIAN\_QUESTION\_EXAMPLE}

Example Clinical Note:  
1: {SENTENCE\_1}  
2: {SENTENCE\_2}  
3: {SENTENCE\_3}

Example Relevant Sentences: [1, 3]

Answer:  
{REFERENCE\_ANSWER}

===== END OF FEW-SHOT =====

Now generate the answer.

Patient Question:  
{PATIENT\_QUESTION}

Clinician Question:  
{CLINICIAN\_QUESTION}

Clinical Note:  
1: {SENTENCE\_1}  
2: {SENTENCE\_2}  
3: {SENTENCE\_3}

Essential Sentences:  
[ID\_LIST]

Final Answer (<=75 words):

# GitAI at ArchEHR-QA 2026: Prompting Strategies and Constitutional AI for Clinical Question Answering

Saran Krishnasamy, Inez Wihardjo

GitAI

San Francisco, CA

saran@gigit.ai, inez@gigit.ai

## Abstract

Answering patient questions from electronic health records requires identifying relevant evidence in lengthy clinical notes and generating faithful, patient-friendly answers. We present a systematic study of LLM prompting strategies for both tasks, evaluating 21 evidence identification methods and 13 answer generation methods across 5 language models. For evidence identification, we find that LLM prompting outperforms traditional retrieval (BM25, SBERT, BioLinkBERT) by 19 F1 points, and that prompt framing alone controls precision/recall trade-offs: inclusive framing achieves 90% recall on dev while balanced framing reaches 67% precision. For answer generation, we apply a Constitutional AI pipeline that critiques and revises answers against five clinical faithfulness principles, improving BLEU and ROUGE over the constrained baseline. Our analysis reveals that chain-of-thought effectiveness is strongly model-dependent, and that simple well-designed prompts outperform complex multi-step pipelines. We evaluate our approaches on the ArchEHR-QA 2026 shared task at CL4Health, achieving 58.0 F1 for evidence identification and 31.8 overall for answer generation.

**Keywords:** clinical NLP, question answering, electronic health records, evidence identification, constitutional AI, large language models

## 1. Introduction

Patients discharged from hospitals frequently have unresolved questions about their diagnoses, treatments, and care plans (Singhal et al., 2023). Electronic health records (EHRs) contain detailed clinical information that could address these questions, but the complexity and volume of clinical notes make them inaccessible to most patients (Thirunavukarasu et al., 2023). Automatically answering patient questions from EHR data requires both identifying relevant evidence within lengthy clinical notes and generating accurate, faithful answers grounded in that evidence.

The ArchEHR-QA 2026 shared task (Soni and Demner-Fushman, 2026b) addresses this challenge through three subtasks: question interpretation (Subtask 1), evidence sentence identification (Subtask 2), and patient-friendly answer generation (Subtask 3). We participate in Subtasks 2 and 3 as team **GitAI**.

Our approach centers on a comprehensive, systematic comparison of methods rather than a single-system submission. For Subtask 2, we evaluate 21 methods across 5 models, spanning traditional retrieval baselines and LLM prompting strategies. For Subtask 3, we evaluate 13 methods including a Constitutional AI pipeline. Our key contributions are:

1. A systematic comparison demonstrating that LLM prompting outperforms retrieval baselines by 19 F1 points for clinical evidence identification.

2. Analysis of prompt framing effects, showing that wording controls precision/recall trade-offs (inclusive:  $R=90.1$ , balanced:  $P=67.2$ ).
3. A principle-guided critique-and-revise pipeline (adapted from Constitutional AI) with five clinical principles for answer generation, improving BLEU by +1.18 and ROUGE-Lsum by +1.67 over the constrained baseline.
4. Analysis of model-prompt interaction effects revealing that chain-of-thought is highly model-dependent (helps GPT-5 but catastrophically hurts Claude Haiku).

## 2. Related Work

### 2.1. Clinical QA from EHRs

Clinical question answering from EHRs has gained attention with datasets such as emrQA (Pampari et al., 2018) and EHRNoteQA (Kweon et al., 2024). The ArchEHR-QA dataset (Soni and Demner-Fushman, 2026a) established benchmarks for multi-turn QA from discharge summaries, with the 2025 shared task (Soni et al., 2025) revealing that LLM-based approaches generally outperformed retrieval-only methods for evidence identification. Recent work has demonstrated that large language models encode substantial clinical knowledge (Singhal et al., 2023), though grounding answers in specific patient records remains challenging.

## 2.2. Evidence Extraction and Retrieval

Traditional retrieval approaches for clinical evidence include sparse methods like BM25 (Robertson and Zaragoza, 2009) and dense retrieval with Sentence-BERT (Reimers and Gurevych, 2019). Domain-specific models such as BioLinkBERT (Yasunaga et al., 2022) and MedCPT (Jin et al., 2023) have shown improvements for biomedical text. Clinical-Longformer (Li et al., 2022) handles the long sequences typical of clinical notes, while cross-encoder reranking (Nogueira and Cho, 2019) with models like BGE (Xiao et al., 2024) provides more precise relevance scoring at higher computational cost.

## 2.3. LLM Prompting Strategies

Chain-of-thought prompting (Wei et al., 2022) has been shown to improve reasoning in LLMs, and self-consistency (Wang et al., 2023) further improves robustness by sampling multiple reasoning paths. DSPy (Khattab et al., 2024) and its MIPROv2 optimizer (Opsahl-Ong et al., 2024) enable automatic prompt optimization. Medical prompt engineering has been explored through approaches like Med-PaLM (Singhal et al., 2023) and Med-prompt (Nori et al., 2023), which combine few-shot ensembling with chain-of-thought for clinical benchmarks. Bogireddy et al. (2025) achieved 2nd place at ArchEHR-QA 2025 using DSPy-optimized prompts with self-consistency voting.

## 2.4. Constitutional AI

Constitutional AI (Bai et al., 2022) was introduced as a training-time method for aligning AI systems to be helpful, harmless, and honest, using self-critique against explicit principles followed by reinforcement learning from AI feedback (RLAIF). While the original framework operates during model training, its core mechanism (evaluating outputs against a set of written principles and revising to resolve violations) can be applied at inference time as a prompting strategy. We adapt this principle-guided critique-and-revise paradigm for clinical answer generation, defining domain-specific principles focused on factual faithfulness, terminology accuracy, and temporal consistency with the source clinical notes. Our approach is also related to self-refinement prompting (Madaan et al., 2023), which iteratively improves outputs using self-generated feedback; we differentiate our pipeline by grounding the critique in domain-specific clinical principles rather than general quality criteria. Our use of the term “Constitutional AI” refers to this inference-time adaptation rather than the full RLAIF training procedure.

## 3. Task Description and Data

### 3.1. Subtask 2: Evidence Identification

Given a patient question, its clinical interpretation, and the full discharge summary segmented into numbered sentences, the task is to identify which sentences constitute essential or supplementary evidence for answering the question. Evaluation uses strict and lenient micro-averaged F1 scores, where lenient scoring gives partial credit for supplementary evidence.

### 3.2. Subtask 3: Answer Generation

Given the same inputs, generate a patient-friendly answer in at most 75 words. Evaluation combines six metrics: BLEU (Papineni et al., 2002), ROUGE-Lsum (Lin, 2004), SARI (Xu et al., 2016), BERTScore (Zhang\* et al., 2020), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023), with the overall score being the average.

### 3.3. Dataset

The ArchEHR-QA 2026 dataset (Soni and Demner-Fushman, 2026a) consists of discharge summaries from MIMIC-IV (Johnson et al., 2024) with patient questions and clinician-authored answers. The data was accessed under a PhysioNet (Goldberger et al., 2000) credentialed data use agreement. Only de-identified discharge summaries as provided by the shared task organizers were used; MIMIC-IV data is already de-identified per HIPAA Safe Harbor standards. Three splits are available: a dev set (20 cases) with gold evidence labels, a test set released from the 2025 shared task (Soni et al., 2025) (100 cases with gold labels, used for additional development), and a new blind test-2026 set (47 cases) for official evaluation. Table 1 summarizes the dataset statistics.

|                         | Dev  | Test | Test-2026 |
|-------------------------|------|------|-----------|
| Number of cases         | 20   | 100  | 47        |
| Avg. sentences/note     | 21.4 | 26.0 | 34.0      |
| Avg. evidence sentences | 7.3  | —    | —         |

Table 1: Dataset statistics for ArchEHR-QA 2026. Test is from the 2025 edition (gold labels available); Test-2026 is the new blind evaluation set.

## 4. System Description

### 4.1. Subtask 2: Evidence Identification

We evaluate 21 methods organized into four categories (Figure 1): non-LLM retrieval baselines, LLM prompting variants, novel methods, and hybrid/ensemble approaches.

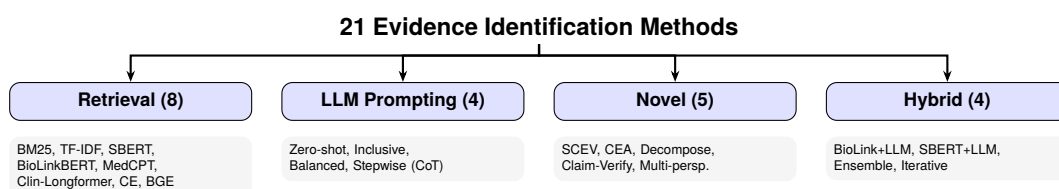


Figure 1: Overview of 21 evidence identification methods organized into four categories.

#### 4.1.1. Non-LLM Retrieval Baselines

We implement eight retrieval baselines using the patient question and clinician interpretation as queries:

**Sparse retrieval:** BM25 (Robertson and Zaragoza, 2009) and TF-IDF (Manning et al., 2008) with cosine similarity.

**Dense retrieval:** Sentence-BERT (all-MiniLM-L6-v2) (Reimers and Gurevych, 2019), BioLinkBERT-large (Yasunaga et al., 2022), MedCPT (Jin et al., 2023), and Clinical-Longformer (Li et al., 2022).

**Reranking:** Cross-encoder reranking (MiniLM (Wang et al., 2020) bi-encoder + cross-encoder) and BGE-large retrieval with BGE-reranker (Xiao et al., 2024).

All retrieval methods return the top- $k$  sentences ( $k=7$  for retrieval,  $k=5$  for reranking). Notably, SBERT, BioLinkBERT, and MedCPT produce identical scores (Table 2); with only  $\sim 21$  sentences per note on average, these dense retrievers return largely overlapping top-7 sets despite different embedding spaces.

#### 4.1.2. LLM Prompting

We design four prompt variants that differ in their framing of the selection task:

**Zero-shot:** A minimal prompt asking the model to identify relevant sentences and output a JSON array.

**Inclusive:** Instructs the model to identify *ALL* relevant sentences, emphasizing “it’s better to include a marginally relevant sentence than to miss important evidence.” This framing optimizes for recall.

**Balanced:** Uses a three-tier categorization (essential, supporting, background) to guide selection granularity, balancing precision and recall.

**Stepwise (CoT):** Decomposes the task into four explicit steps (identify key concepts, locate relevant sentences, find supporting context, and compile the final list), eliciting chain-of-thought reasoning (Wei et al., 2022).

#### 4.1.3. Novel Methods

**SCEV (Self-Consistency Evidence Verification):** Inspired by self-consistency (Wang et al., 2023), we generate 5 evidence selections using the balanced prompt at varying temperatures (0.3 to 0.9),

then apply majority voting with a 40% threshold. Sentences selected by at least 2 of 5 samples are retained.

**CEA (Counterfactual Evidence Attribution):** We first obtain a broad candidate set using inclusive prompting. For each candidate sentence, we test its importance by removing it from the note and asking whether the question can still be fully answered. Sentences whose removal renders the question unanswerable are classified as essential.

**Question Decomposition:** We decompose the patient question into sub-questions and independently identify evidence for each, then take the union.

**Claim Verification:** We extract claims from the question, then verify each claim against individual sentences to find supporting evidence.

**Multi-Perspective:** We prompt from three perspectives (clinical accuracy, completeness, patient relevance) and combine selections via union.

#### 4.1.4. Hybrid and Ensemble Approaches

**BioLinkBERT + LLM:** Pre-filter to top-15 with BioLinkBERT, then use LLM to refine.

**SBERT + LLM:** Pre-filter to top-15 with SBERT, then use LLM to refine from this candidate set.

**Ensemble Voting:** Run three prompts (zero-shot, inclusive, balanced) and retain sentences selected by at least one method.

**Iterative Refinement:** Generate initial selection, self-critique for missed evidence and false positives, then produce a revised selection.

### 4.2. Subtask 3: Answer Generation

We evaluate 13 methods spanning direct prompting, multi-stage pipelines, a Constitutional AI approach, and DSPy-optimized modules.

#### 4.2.1. Direct Prompting Methods

**Constrained:** The required baseline that instructs the model to answer using *only* information from the clinical notes, with a 75-word limit and five explicit constraints (no external knowledge, professional terminology, acknowledge limitations).

**Zero-shot:** Minimal prompt with just the question and notes.



**Balanced:** Adds the clinician’s interpretation and structured instructions for balanced coverage.

#### 4.2.2. Multi-Stage Methods

**CoT-Verified:** Two-stage pipeline: (1) chain-of-thought generation with explicit identify, locate, synthesize, verify steps, followed by (2) self-verification against the notes.

**Evidence-First:** Leverages Subtask 2 output by first extracting evidence sentences, then generating the answer from only the identified evidence.

**Iterative Refinement:** Three-stage generate, critique, revise pipeline using separate prompts for each stage.

**Multi-Perspective:** Elicits responses from three simulated experts (conservative, comprehensive, patient-centered), then synthesizes a final answer.

**Claim-Verified:** Extracts individual claims with evidence citations, then synthesizes verified claims into a coherent answer.

**Decompose-Synthesize:** Decomposes the question into 2 to 4 sub-questions, answers each independently from the notes, then synthesizes a unified response.

#### 4.2.3. Constitutional AI Pipeline

Our approach adapts the principle-guided critique-and-revise mechanism from Constitutional AI (Bai et al., 2022) as an inference-time prompting strategy for clinical faithfulness. The pipeline has three stages (Figure 2):

**Stage 1 (Generate):** Produce an initial answer using the constrained prompt.

**Stage 2 (Critique):** Evaluate the draft against five clinical principles:

1. All claims must be directly traceable to the clinical notes
2. Do not speculate or make inferences beyond documentation
3. Acknowledge when information is incomplete
4. Use exact medical terminology from the notes
5. Maintain temporal accuracy (do not confuse past/present)

The model evaluates each principle as PASS or FAIL with explanations.

**Stage 3 (Revise):** Generate a revised answer that addresses all principle violations while maintaining the 75-word constraint.

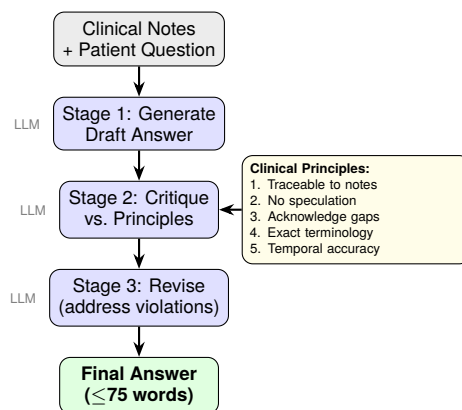


Figure 2: Constitutional AI pipeline for clinical answer generation. Each stage uses a separate LLM call. The five principles enforce clinical faithfulness during the critique stage.

#### 4.2.4. DSPy Optimization

Following the success of DSPy at ArchEHR-QA 2025 (Bogireddy et al., 2025), we implement three DSPy modules optimized with MIPROv2 (Opsahl-Ong et al., 2024):

**DirectQA:** Single-stage ChainOfThought answer generation.

**TwoStageQA:** First extracts evidence sentence IDs, then generates from the extracted evidence.

**SelfConsistencyQA:** Generates 5 candidate answers and selects the most consistent one.

All modules are optimized on the 20-example dev set using MIPROv2 with the “light” auto-configuration. The optimization metric combines ROUGE-Lsum similarity between the generated and gold answers with a penalty for exceeding the 75-word limit.

## 5. Experimental Setup

### 5.1. Models

We evaluate across models from two providers:

**OpenAI (API):** GPT-5, GPT-4o, and GPT-5.2 (an updated GPT-5 variant with improved instruction following) (OpenAI, 2025).

**Anthropic (AWS Bedrock):** Claude Opus 4.5 and Claude Haiku 4.5 (Anthropic, 2025).

Not all models are evaluated on all methods due to API cost constraints; Tables 3 and 4 report the most informative subset. Opus was evaluated only on zero-shot for Subtask 2.

### 5.2. Hyperparameters

For Subtask 2, LLM methods use temperature 0.0 for deterministic output. Retrieval methods use top- $k=7$  (retrieval) or  $k=5$  (reranking). SCEV uses 5 samples at temperatures 0.3 to 0.9 with a 40%

voting threshold. For Subtask 3, generation uses temperature 0.3 with max\_tokens = 500, and all outputs are truncated to 75 words.

### 5.3. Submission

For Subtask 2, we submit the inclusive prompt with GPT-5, prioritizing recall. We selected lenient F1 over strict F1 as the evaluation criterion because it assigns partial credit for supplementary evidence, better reflecting the downstream utility of evidence retrieval: for answer generation (Subtask 3), including supplementary context sentences is preferable to missing them entirely. Although other prompts achieve higher lenient F1 (e.g., stepwise: 69.0), the inclusive prompt’s near-complete recall (90.1% vs. 64.5% for balanced) minimizes the risk of missing evidence that benefits Subtask 3. For Subtask 3, we submit the constitutional method with Claude Opus 4.5 based on the best dev overall score.

## 6. Results and Analysis

### 6.1. Subtask 2: Evidence Identification

#### 6.1.1. Dev Set Results

Table 2 shows the dev set results comparing non-LLM baselines and LLM methods. LLM prompting methods substantially outperform retrieval baselines. The best retrieval method (SBERT/BioLinkBERT, 46.7 strict F1) is surpassed by the best LLM prompting (stepwise GPT-5, 65.9 strict F1), a gap of 19.2 points. Even the simplest zero-shot prompt (64.2) outperforms all retrieval baselines.

#### 6.1.2. Cross-Model Comparison

Table 3 shows how prompt strategies interact with different models. The stepwise (CoT) prompt shows the most dramatic model-dependency: it achieves 65.9 F1 with GPT-5 but only 16.7 with Claude Haiku 4.5, a catastrophic 49-point drop. This is because Haiku, a smaller model, struggles with the multi-step reasoning format, often failing to produce valid JSON output after the reasoning chain.

#### 6.1.3. Test Results

Our test submission uses inclusive prompting with GPT-5, achieving **58.0 lenient micro F1** (rank 12). As discussed in Section 5.3, we prioritized recall (90.1% on dev) over F1 to maximize evidence coverage for downstream answer generation, accepting the trade-off of lower precision (45.4%).

| Method                               | Model | S-F1        | S-P         | S-R         |
|--------------------------------------|-------|-------------|-------------|-------------|
| <i>Non-LLM Retrieval Baselines</i>   |       |             |             |             |
| BM25                                 | –     | 39.1        | 36.4        | 42.1        |
| TF-IDF                               | –     | 41.4        | 38.6        | 44.6        |
| SBERT                                | –     | 46.7        | 43.6        | 50.4        |
| BioLinkBERT                          | –     | 46.7        | 43.6        | 50.4        |
| MedCPT                               | –     | 46.7        | 43.6        | 50.4        |
| Clin-Longformer                      | –     | 40.6        | 37.9        | 43.8        |
| MiniLM+CE                            | –     | 38.0        | 42.0        | 34.7        |
| BGE+Rerank                           | –     | 43.4        | 48.0        | 39.7        |
| <i>LLM Prompting Methods (GPT-5)</i> |       |             |             |             |
| Zero-shot                            | GPT-5 | 64.2        | 63.9        | 64.5        |
| Inclusive                            | GPT-5 | 60.4        | 45.4        | <b>90.1</b> |
| Balanced                             | GPT-5 | 65.8        | <b>67.2</b> | 64.5        |
| Stepwise (CoT)                       | GPT-5 | <b>65.9</b> | 59.2        | 74.4        |
| <i>Novel Methods (GPT-5)</i>         |       |             |             |             |
| SCEV                                 | GPT-5 | 64.4        | 60.0        | 69.4        |
| CEA                                  | GPT-5 | 46.5        | 46.0        | 47.1        |
| Decomposition                        | GPT-5 | 42.3        | 39.6        | 45.5        |
| Claim Verify                         | GPT-5 | 50.9        | 57.9        | 45.5        |
| Multi-persp.                         | GPT-5 | 60.6        | 54.2        | 68.6        |
| <i>Hybrid / Ensemble (GPT-5)</i>     |       |             |             |             |
| BioLink+LLM                          | GPT-5 | 55.5        | 54.8        | 56.2        |
| SBERT+LLM                            | GPT-5 | 53.8        | 69.7        | 43.8        |
| Ensemble                             | GPT-5 | 58.1        | 41.5        | 96.7        |
| Iterative                            | GPT-5 | 58.4        | 48.1        | 74.4        |

Table 2: Subtask 2 dev results (strict micro scores). S-F1 = strict micro F1, S-P = precision, S-R = recall. Best per column in **bold**.

| Prompt    | GPT-5       | 5.2  | 4o   | Opus | Haiku |
|-----------|-------------|------|------|------|-------|
| Zero-shot | <b>64.2</b> | 53.2 | 49.8 | 55.5 | 52.6  |
| Inclusive | 60.4        | 58.5 | 59.5 | –    | 60.8  |
| Balanced  | 65.8        | 59.9 | 56.4 | –    | 57.9  |
| Stepwise  | 65.9        | 59.4 | 53.2 | –    | 16.7  |
| SCEV      | 64.4        | 60.2 | 58.9 | –    | 59.3  |

Table 3: Subtask 2 cross-model comparison (strict micro F1). Opus = Claude Opus 4.5, Haiku = Claude Haiku 4.5. GPT-5 consistently leads.

### 6.2. Subtask 3: Answer Generation

#### 6.2.1. Dev Set Results

Table 4 shows the dev set results for Subtask 3. The Constitutional AI method with Claude Opus 4.5 achieves the highest overall score (31.4), though the margin over the constrained baseline (30.7) is modest given the small dev set size (20 cases), and the differences may not be statistically significant. The improvement is primarily driven by BLEU (+1.18) and ROUGE-Lsum (+1.67). Notably, BERTScore is remarkably stable across methods (34.6 to 40.3), while BLEU varies from 1.77 to 6.29, suggesting that BERTScore is less discriminative for this task and that overall score differences are driven primarily by lexical overlap metrics. Note that

AlignScore and MEDCON were not computed during dev evaluation (both zero for all methods), so the dev overall score averages only BLEU, ROUGE-Lsum, SARI, and BERTScore.

### 6.2.2. Test Results

Our test submission uses the constitutional method with Claude Opus 4.5, achieving an overall score of **31.8** (rank 10).

## 6.3. Analysis

### 6.3.1. LLMs vs. Retrieval for Evidence Identification

The most striking finding from Subtask 2 is the large gap between retrieval baselines and LLM prompting. The best retrieval method achieves 46.7 strict F1, while the best LLM prompt (stepwise GPT-5) reaches 65.9, a gap of 19.2 points. This suggests that LLMs can effectively perform the sentence-level relevance assessment required for clinical evidence identification, leveraging their understanding of clinical context in ways that embedding similarity alone cannot capture.

#### 6.3.2. Prompt Framing and Precision/Recall

Our four prompt variants reveal a clear precision/recall spectrum controlled entirely by framing:

- **Inclusive** (R=90.1, P=45.4): Maximizes recall by instructing “be INCLUSIVE.”
- **Stepwise/CoT** (R=74.4, P=59.2): Structured reasoning improves precision while maintaining high recall.
- **Zero-shot** (R=64.5, P=63.9): Balanced default behavior.
- **Balanced** (R=64.5, P=67.2): Three-tier categorization yields the highest precision.

This finding has practical implications: practitioners can tune the precision/recall trade-off purely through prompt design, without any model retraining. In clinical settings, inclusive framing is preferred when missed evidence carries high risk (e.g., safety-critical questions about medications or procedures), while balanced framing suits scenarios where false positives are costly (e.g., automated summarization where irrelevant content degrades output quality).

#### 6.3.3. Chain-of-Thought is Model-Dependent

The stepwise (CoT) prompt exhibits dramatic model-dependency. With GPT-5, it achieves the highest F1 (65.9) of any single prompt. With Claude Haiku

4.5, it produces only 16.7 F1, a catastrophic failure. Analysis of Haiku’s outputs reveals that the smaller model often fails to complete the reasoning chain, producing malformed JSON or getting stuck in verbose reasoning without reaching the final answer format. Intermediate models (GPT-5.2, GPT-4o) show modest benefits from CoT, suggesting that the effectiveness of structured reasoning scales with model capability. Practitioners should be cautious with CoT prompting for smaller models; in our experiments, Claude Haiku catastrophically fails at structured reasoning, while GPT-4o handles it adequately but with degraded performance relative to simpler prompts.

To illustrate, consider Case 4 (a patient asking about cardiac catheterization for heart failure; 7 essential and 2 supplementary gold evidence sentences). The inclusive prompt with GPT-5 retrieves 12 sentences, capturing 8 of 9 gold sentences (missing only sentence 10). In contrast, the stepwise prompt returns only sentence 5: the multi-step reasoning caused the model to over-filter, retaining just one of nine relevant sentences. While stepwise achieves the best overall F1 across the dev set, its performance is highly variable per case; Case 4 represents a worst-case failure mode where chain-of-thought over-filtered. See Appendix B for the full comparison.

These results suggest practical conditions for CoT reliability: structured reasoning prompts should be reserved for high-capability models (here, GPT-5 class or above) that can maintain coherent multi-step reasoning while adhering to output format constraints. For smaller or less capable models, the added complexity of CoT increases the risk of format violations (e.g., malformed JSON) and reasoning failures that degrade performance below simpler prompts. Practitioners should validate CoT effectiveness on their target model before deployment rather than assuming transferability across model scales.

#### 6.3.4. Principle-Guided Critique Improves Answer Quality

The Constitutional AI pipeline achieves the best overall score on both dev (31.4) and test (31.8) for Subtask 3, though the dev margin (31.4 vs. 30.7) is small and may not be significant given only 20 evaluation cases. Comparing constitutional to constrained (same model, Claude Opus 4.5), the critique-and-revise cycle improves BLEU by +1.18 (5.11 → 6.29) and ROUGE-Lsum by +1.67 (19.8 → 21.5), while SARI remains stable (58.5).

Analysis of principle violations across the 20 dev cases reveals that Principle 1 (traceability to notes) is the most frequently violated, failing in 9 of 20 drafts (45%), followed by Principle 2 (no speculation) which often co-occurs with Principle 1 viola-

| Method                   | Model    | Ovr.        | BLEU        | R-Lsum      | SARI        | BERTSc.     |
|--------------------------|----------|-------------|-------------|-------------|-------------|-------------|
| Constitutional           | Opus 4.5 | <b>31.4</b> | <b>6.29</b> | <b>21.5</b> | <b>58.5</b> | 39.2        |
| Constrained              | Opus 4.5 | 30.7        | 5.11        | 19.8        | 58.5        | 39.3        |
| DSPy-SC                  | GPT-4o   | 30.6        | 5.23        | 21.4        | 56.9        | 38.8        |
| DSPy-Direct              | GPT-4o   | 30.5        | 4.62        | 20.9        | 56.2        | <b>40.3</b> |
| DSPy-SC                  | GPT-5.2  | 30.4        | 4.42        | 21.5        | 56.3        | 39.4        |
| Constrained <sup>†</sup> | Opus 4.5 | 30.4        | 4.55        | 19.8        | 58.3        | 38.9        |
| DSPy-Direct              | GPT-5.2  | 29.9        | 4.07        | 20.7        | 57.1        | 37.9        |
| DSPy-SC                  | Opus 4.5 | 29.0        | 2.39        | 19.0        | 55.6        | 39.1        |
| Iterative                | Opus 4.5 | 27.9        | 2.53        | 18.2        | 55.3        | 35.5        |
| DSPy-2Stage              | GPT-4o   | 26.7        | 1.77        | 18.7        | 51.9        | 34.6        |

Table 4: Answer generation dev results sorted by overall score. R-Lsum = ROUGE-Lsum, BERTSc. = BERTScore. Overall averages BLEU, R-Lsum, SARI, and BERTScore (AlignScore and MEDCON were zero for all methods during dev evaluation). <sup>†</sup>Earlier run with different random seed. DSPy-SC = SelfConsistency.

tions. Principles 3 to 5 are rarely triggered. Common error patterns include adding unsupported severity qualifiers (“severe,” “significant”), speculating about causal mechanisms not documented in the notes, and making inferences beyond what is explicitly stated. Table 5 illustrates how the critique stage identifies these violations and the revision replaces speculation with documented facts.

However, in cases where the draft already passes all principles (11 of 20 dev cases), the critique-and-revise cycle adds computational cost without meaningful improvement. Additionally, the 75-word constraint means revisions sometimes drop relevant information to accommodate corrections, trading completeness for faithfulness. We discuss these cost trade-offs further in the Limitations section.

### 6.3.5. DSPy Optimization Performance

DSPy-optimized modules with MIPROv2 produce competitive results (30.5 to 30.6 overall) but do not meaningfully improve over the constrained baseline (30.7 with Claude Opus 4.5). Given the small dev set (20 examples), the differences among DSPy, constrained, and constitutional methods are within the noise margin and likely not statistically significant. This suggests that automatic prompt optimization with MIPROv2 on a 20-example dev set provides insufficient signal to outperform a well-designed manual prompt for this task. The Self-ConsistencyQA module slightly outperforms DirectQA (30.6 vs. 30.5 with GPT-4o), consistent with the self-consistency literature (Wang et al., 2023). The TwoStageQA module underperforms (26.7), suggesting that the two-stage pipeline introduces additional complexity without corresponding gains at this scale.

## 7. Discussion

### 7.1. Practical Implications

Our systematic comparison provides practical guidance for clinical QA system design. For evidence identification, simple LLM prompting with clear framing outperforms complex retrieval pipelines, hybrid systems, and novel multi-step approaches. The inclusive prompt’s high recall makes it suitable when downstream tasks (e.g., answer generation) benefit from comprehensive evidence, while balanced prompting is preferable for precision-critical applications.

For answer generation, the Constitutional AI approach demonstrates that domain-specific self-critique can improve answer quality at modest computational cost (3 LLM calls vs. 1). The five clinical principles provide an interpretable framework that could be extended with additional domain expertise.

To guide practitioners in selecting a prompt strategy: inclusive framing (high recall) is recommended when downstream tasks depend on comprehensive evidence coverage, for example, when generating patient-facing answers where missing a relevant clinical detail could lead to an incomplete or misleading response. Balanced framing (high precision) is preferable in precision-critical settings such as clinical decision support, where false positives could waste clinician attention or introduce noise into automated workflows. The zero-shot prompt provides a reasonable default when no strong prior exists about the precision/recall requirements of the application.

### 7.2. Dev/Test Gap

The constitutional method scores 31.4 on dev and 31.8 on test, though these are not directly comparable: the dev overall averages 4 metrics (AlignScore and MEDCON were not computed), while the test



| Case 2: Why was father given Lasix?                                                                                                                                                                                                                                                                | Case 4: Why cardiac catheterization?                                                                                                                                                                                                         | Case 7: Why was Coumadin stopped?                                                                                                                                                                                                              |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Draft</b> (red = unsupported claims)</p> <p>“...acute diastolic heart failure with <b>significant</b> volume overload...and <b>severe</b> lower extremity edema. [...] His oxygen was <b>likely reduced as diuresis improved his respiratory status</b>.”</p>                                | <p>“...revealed elevated filling pressures...and reduced cardiac output (Fick cardiac index of 1.72), <b>confirming severe low-output heart failure</b>. The milrinone trial during catheterization was successful...”</p>                   | <p>“...Coumadin was discontinued following your subarachnoid hemorrhage and IVC filter placement. [...] <b>The discontinuation is directly related to your brain hemorrhage, as anticoagulation poses bleeding risks in this context</b>.”</p> |
| <p><b>Critique</b> (principle violations identified)</p> <p>P1 FAIL: “severe” not in notes; causal relationship “likely reduced as diuresis improved” not stated</p>                                                                                                                               | <p>P1 FAIL: notes say “low-output” without the qualifier “severe”; P2 FAIL: “confirming” implies diagnostic certainty not stated</p>                                                                                                         | <p>P1 FAIL: reason for discontinuation not explicitly stated; P2 FAIL: “anticoagulation poses bleeding risks” is external medical reasoning</p>                                                                                                |
| <p><b>Revision</b> (teal = corrected/added from notes)</p> <p>“...acute diastolic heart failure with volume overload...and lower extremity edema. The Lasix was administered intravenously, including via continuous drip, <b>then transitioned to oral torsemide</b> to remove excess fluid.”</p> | <p>“...showed elevated filling pressures (mean PCW of 30) and reduced cardiac output (Fick C.I. of 1.72), <b>consistent with</b> low output heart failure. The milrinone trial was successful, <b>improving your Fick C.I. to 2.79</b>.”</p> | <p>“...Coumadin was discontinued following your IVC filter placement. [...] <b>The clinical notes do not explicitly state the reason for discontinuation</b> beyond documenting it occurred after filter placement.”</p>                       |

Table 5: Constitutional AI examples across three cases. **Red text** in drafts marks unsupported claims flagged by the critique. **Teal text** in revisions highlights corrections: added documented details, softened language, or explicit acknowledgment of information gaps.

overall averages all 6 metrics. For Subtask 2, the dev/test gap is larger: the inclusive prompt achieves 64.3 lenient micro F1 on dev (60.4 strict) but 58.0 lenient on test. This may reflect the test set’s greater diversity (47 vs. 20 cases) or distributional shift in question types and clinical note complexity.

## 8. Conclusion

We presented GigitAI’s systems for ArchEHR-QA 2026 Subtasks 2 and 3, featuring a systematic comparison of 21 evidence identification methods and 13 answer generation methods across 5 language models. Our key findings are that LLM prompting outperforms traditional retrieval by a wide margin (19 F1 points) for clinical evidence identification, prompt framing provides fine-grained control over precision/recall trade-offs, chain-of-thought effectiveness is strongly model-dependent, and Constitutional AI with clinical principles can improve answer generation quality by catching unsupported claims in LLM drafts. Our best submissions achieve 58.0 F1 (rank 12) for evidence identification and 31.8 overall (rank 10) for answer generation.

## 9. Limitations

Our study has several limitations. First, the small dev set (20 cases) limits our ability to draw statistically robust conclusions; even the same method

with different random seeds varies by 0.3 points (Table 4), illustrating the measurement noise at this scale. Second, we rely entirely on prompting and do not explore fine-tuning, which could improve performance, particularly for smaller models where CoT prompting fails. Third, our multi-stage pipelines (Constitutional AI, CoT-Verified, Iterative Refinement) require 2 to 3 LLM calls per answer, multiplying both latency and cost relative to single-call methods. For the Constitutional AI pipeline specifically, the critique-and-revise cycle adds no value in 55% of cases (11/20 dev cases pass all principles on the first draft), making the additional cost wasteful for already-faithful outputs. A confidence-based gating mechanism could skip the critique stage when the draft is likely faithful, reducing average cost. Fourth, we rely entirely on proprietary LLM APIs (OpenAI, Anthropic), which limits reproducibility and raises concerns about cost, data privacy, and access continuity. Evaluating open-weight models (e.g., Llama (Grattafiori et al., 2024), Mistral (Jiang et al., 2023)) would strengthen the generalizability of our findings. Fifth, the 75-word constraint for Subtask 3 makes it difficult to produce comprehensive answers for complex clinical questions. Finally, our clinical principles are manually designed; automatic principle discovery could further improve the approach.



## 10. Bibliographical References

- Anthropic. 2025. [Claude opus 4.5 and claude haiku 4.5 system cards](#). Technical report, Anthropic.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanhham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. [Constitutional AI: Harmlessness from AI Feedback](#). ArXiv:2212.08073 [cs].
- Sai Prasanna Teja Reddy Bogireddy, Abrar Ma-jeedi, Viswanath Gajjala, Zhuoyan Xu, Siddhant Rai, and Vaishnav Potlapalli. 2025. [Neural at ArchEHR-QA 2025: Agentic prompt optimization for evidence-grounded clinical question answering](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 104–109, Vienna, Austria. Association for Computational Linguistics.
- Ary Goldberger, Luís Amaral, Leon Glass, Jeffrey Hausdorff, Plamen Ivanov, Roger Mark, Joseph Mietus, George Moody, Chung-Kang Peng, and H. Stanley. 2000. [Physiobank, physiokit, and physionet : Components of a new research resource for complex physiologic signals](#). *Circulation*, 101:E215–20.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- 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.
- Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. [MIMIC-IV](#). *PhysioNet*. Version 3.1.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. [Dspy: Compiling declarative language model calls into self-improving pipelines](#). In *The Twelfth International Conference on Learning Representations*.
- Sunjun Kweon, Jiyou Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwang Hyun Kim, Seunghyun Won, and Edward Choi. 2024. [EHRNoteQA: A Patient-Specific Question Answering Benchmark for Evaluating Large Language Models in Clinical Settings](#). *PhysioNet*. Version 1.0.0.
- Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. [Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences](#). *CoRR*, abs/2201.11838.
- 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.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, 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.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Sch  tze. 2008. [Introduction to Information Retrieval](#). Cambridge University Press, Cambridge, UK.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. [Passage re-ranking with bert](#). *arXiv preprint arXiv:1901.04085*.
- Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas

- King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. [Can generalist foundation models out-compete special-purpose tuning? case study in medicine](#).
- OpenAI. 2025. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*.
- Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. [Optimizing instructions and demonstrations for multi-stage language model programs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- 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 on Association for Computational Linguistics, ACL '02*, page 311–318, USA. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentencebert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.
- 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, Blaise Agüera Y Arcas, Dale Webster, Greg S Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkumar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. [Large language models encode clinical knowledge](#). *Nature*, 620(7972):172—180.
- Sarvesh Soni and Dina Demner-Fushman. 2026a. [A dataset for addressing patient's information needs related to clinical course of hospitalization](#). *Scientific Data*.
- 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.
- 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, Viena, Austria. Association for Computational Linguistics.
- Arun Thirunavukarasu, Darren Ting, Kabilan Elangovan, Laura Gutierrez Sinisterra, Ting Tan, and Daniel Ting. 2023. [Large language models in medicine](#). *Nature Medicine*, 29.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20*, Red Hook, NY, USA. Curran Associates Inc.
- 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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- 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.
- 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, SIGIR '24*, page 641–649, New York, NY, USA. Association for Computing Machinery.

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.

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.

Wen-wai Yim, Yajuan Velvin 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.

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*.

## A. Prompt Templates

This appendix lists the key prompt templates used in our systems. Placeholders are shown in {braces}.

### A.1. Evidence Identification (Subtask 2)

#### Inclusive Prompt:

*You are a clinical expert identifying relevant evidence from medical records.*

*Patient Question:* {patient\_question}  
*Clinician's Interpretation:* {clinical\_question}

*Clinical Note Sentences:* {numbered\_sentences}

*Your task: Identify ALL sentences that contain information relevant to answering the question.*

*IMPORTANT: Be INCLUSIVE. It's better to include a marginally relevant sentence than to miss important evidence.*

*Output a JSON array of ALL relevant sentence numbers.*

#### Balanced Prompt:

*Identify sentences that contain information needed to answer the question.*

*Selection criteria: 1. ESSENTIAL: Sentences that directly answer part of the question (MUST include) 2. SUPPORTING: Sentences that provide important context or values (SHOULD include) 3. BACKGROUND: Sentences that help understand the clinical situation (MAY include)*

#### Stepwise (CoT) Prompt:

*Let's identify evidence step by step: Step 1: What are the key clinical concepts in the question? Step 2: Which sentences contain information about these concepts? Step 3: Which sentences provide supporting context or values? Step 4: Compile the final list of relevant sentences.*

*Think through each step, then output ONLY a JSON array of sentence numbers.*

### A.2. Answer Generation (Subtask 3)

#### Constrained Prompt:

*You are a clinical expert answering a patient's question about their hospitalization.*

*Patient Question:* {patient\_question}  
*Clinician's Interpretation:* {clinical\_question}

*Clinical Notes:* {note\_excerpt}

*Instructions: 1. Answer ONLY using information from the clinical notes above 2. Use professional medical terminology (not overly simplified) 3. If information is incomplete, acknowledge limitations 4. Maximum 75 words, approximately 5 sentences 5. Do NOT speculate or add external medical knowledge*

#### Constitutional AI, Critique Prompt:

*Clinical Principles: 1. All claims must be directly traceable to the clinical notes 2. Do not speculate or make inferences beyond documentation 3. Acknowledge when information is incomplete 4. Use exact medical terminology from the notes 5. Maintain temporal accuracy (don't confuse past/present)*

*Review this answer against the clinical principles above.*

*Clinical Notes:* {note\_excerpt}

*Draft Answer:* {draft}

*For each principle, evaluate compliance (PASS/FAIL) and explain:*

#### Constitutional AI, Revision Prompt:

*Revise this answer to comply with all clinical principles.*

*Original Answer:* {draft}

*Principle Evaluation:* {critique}

*Clinical Notes (for reference):* {note\_excerpt}

*Revised answer (must comply with ALL principles, under 75 words):*

## B. Evidence Identification Example

Table 6 shows a detailed comparison of evidence identification methods on Case 4, where the patient asks “*Why was cardiac catheterization recommended?*”. Gold evidence includes 7 essential and 2 supplementary sentences across the discharge summary. We compare four methods: BM25 (sparse retrieval), SBERT (dense retrieval), inclusive prompt (GPT-5), and stepwise/CoT prompt (GPT-5).

The inclusive prompt captures 8 of 9 gold sentences (missing only sentence 10), while stepwise/CoT retrieves only 1 sentence, demonstrating the over-filtering failure mode discussed in Section 6.3. BM25 captures 4 of 9 gold sentences and SBERT captures 5 of 9, but they select different subsets: BM25 favors lexically similar sentences (e.g., “milrinone”), while SBERT favors semantically similar ones (e.g., “heart failure”).

| ID                                           | Gold  | Sentence (truncated)                                                                                   | BM25 | SBERT | Inclusive  | Stepwise |
|----------------------------------------------|-------|--------------------------------------------------------------------------------------------------------|------|-------|------------|----------|
| 2                                            |       | Cardiology service: abdominal pain, nausea, vomiting felt to be secondary to congestive hepatopathy    | ✓    |       | ✓          |          |
| 4                                            |       | He was aggressively diuresed with a net 10 liters negative since admission                             | ✓    |       |            |          |
| 5                                            | Ess.  | He underwent RHC for milrinone trial, which proved to be successful                                    | ×    | ×     | ✓          | ✓        |
| 6                                            |       | His mean PCW went from 30 to 22, and his Fick C.I. went from 1.72 to 2.79                              |      |       | ✓          |          |
| 7                                            | Supp. | Brief Hospital Course: 48M with idiopathic dilated cardiomyopathy (EF 25%...)                          | ×    | ✓     | ✓          | ×        |
| 8                                            |       | Acute-on-chronic systolic heart failure                                                                |      | ✓     | ✓          |          |
| 9                                            | Supp. | Underlying exacerbation of chronic heart failure, known idiopathic cardiomyopathy...                   | ×    | ×     | ✓          | ×        |
| 10                                           | Ess.  | Last echo showed LVEF = 25%                                                                            | ×    | ×     | ×          | ×        |
| 11                                           | Ess.  | Admitted in the setting of low output heart failure, leading to increased intra-abdominal pressures... | ✓    | ×     | ✓          | ×        |
| 13                                           | Ess.  | Cardiac output and wedge pressure significantly improved after milrinone infusion                      | ✓    | ✓     | ✓          | ×        |
| 14                                           |       | In the ICU maintained on milrinone at 0.5 mcg/hr, transferred to floor                                 | ✓    |       |            |          |
| 16                                           |       | Heart failure specialists had honest discussions about long-term prognosis                             |      | ✓     | ✓          |          |
| 18                                           | Ess.  | You were admitted to the hospital with worsening heart failure                                         | ✓    | ✓     | ✓          | ×        |
| 19                                           | Ess.  | You had a cardiac catheterization that showed you would benefit from milrinone                         | ✓    | ✓     | ✓          | ×        |
| 20                                           | Ess.  | You were started on a milrinone drip, with improvement in your heart's pump function                   | ×    | ✓     | ✓          | ×        |
| <b>Gold sentences found / total gold (9)</b> |       |                                                                                                        | 4/9  | 5/9   | <b>8/9</b> | 1/9      |
| <b>False positives</b>                       |       |                                                                                                        | 3    | 2     | 4          | 0        |

Table 6: Evidence identification comparison on Case 4 (“*Why was cardiac catheterization recommended?*”). Shaded rows are gold evidence (Ess. = essential, Supp. = supplementary). ✓ = correctly identified gold sentence; × = missed gold sentence; ✓ = selected but not gold (false positive). Only sentences selected by at least one method are shown.





# Author Index

- Abubakar, Joseph Itopa, 344  
Adewunmi, Mary Adetutu, 344  
Adlberger, Selina, 117  
Akhtar, Md. Shad, 320  
AL-Smadi, Mohammad, 333  
AlMannaa, Feras, 539  
Amorim, Evelin, 50  
Ananiadou, Sophia, 190  
Anglés-Herrero, Emma, 159  
Aramaki, Eiji, 107, 149  
Arvan, Mohammad, 487
- Bauernfeind, Sophie, 96, 117  
Bel-Enguix, Gemma, 159  
Bernad, Jordi, 159  
Beynen, Pieter, 15  
Bheda, Nishit, 1  
Boekesteijn, Otto, 15  
Bogireddy, Sai Prasanna Teja Reddy, 530  
Bojic, Iva, 26  
Boulogeorgou, Artemis, 190  
Büns, Jan-Henning, 455
- Carroll, Lucien, 428  
Cenikj, Gjorgjina, 223  
Chandramohan, Anuradha, 268  
Chawla, Chirag, 320  
Chen, Lujie Karen, 506  
Chen, Wei-Chun, 362  
Chen, Yu-Xuan, 362  
Choi, Edward, 234  
Choppa, Tejaswi, 469  
Christiansen, Alexander, 441  
Chu, Bohao, 455  
Chung, I-Fang, 362  
Corona Beomont, Ángel Óscar, 159  
Cosma, Adrian, 132
- Damm, Hendrik, 455  
de Boer, Maaïke, 171  
Demner-Fushman, Dina, 255, 296  
di Buono, Maria Pia, 182  
Dias, Gaël, 384  
Díaz Galiano, Manuel Carlos, 72
- Dumitrache, Cosmin, 132  
Džeroski, Sašo, 223
- Eapen, Anu, 268  
Eftimov, Tome, 223  
Ekanayake, Vinu, 207
- Fain, Mary Ellen M., 307  
Faruque, Omar, 506  
Feinstein, Rebecca, 487  
Ferdowsi, Sohrab, 234  
Fernandez-Quilez, Alvaro, 61  
Ferrazzi, Pietro, 245  
Fodeh, Samah, 515  
Frew, Michael, 1  
Friedrich, Christoph M., 455  
Fu, Sunyang, 551  
Fuhr, Norbert, 455
- Gajjala, Viswanatha Reddy, 530  
Galušćáková, Petra, 61  
Garrido, Rúben, 441  
Gayen, Soumya, 296  
Ghosh, Soumitra, 245  
Gjorgjevikj, Ana, 223  
Guglani, Lokesh, 307  
Gupta, Deepak, 296
- H, Goldwin, 268  
Haeri, Hossein, 487  
Hamnett, Leon, 344  
Han, Lifeng, 82  
Hayakawa, Masayo, 107  
Hêche, Félicien, 234  
Hidayah, Rohmah, 149  
Ho, Andy Hau Yan, 26  
Horn, Peter A., 455  
Hosen, Md Biplob, 506  
Huang, Ruihong, 551  
Hussein, Md Alomgeer, 506
- Ide, Hiroo, 107  
Ieva, Francesca, 412  
Igwezeke, Favour, 344  
Ilchenko, Igor, 41

Ionov, Maxim, 159  
Irankhah, Elyas, 515  
Irastortza-Urbietza, Xabier, 434  
Iturra Bocaz, Gabriel, 61

Jansen, Joris, 15  
Jonker, Richard A. A., 441  
Jurowetzki, Roman, 441

Kaponi, Fotini, 190  
Karl, Fabian Tobias, 469  
Kavuluru, Ramakanth, 207  
Khong, Andy, 26  
Kim, Siyoon, 369  
Kolmogorova, Anastasia, 41  
Koroušič Seljak, Barbara, 223  
Krauss, Oliver, 96, 117  
Krishnamurthy, Parameswari, 268  
Krishnasamy, Saran, 565  
Kumari, Jyoti, 395  
Kuramoto, Mana, 107

Lavelli, Alberto, 245  
Lawate, Ashwini Yuvraj, 26  
Lee, Jooyeon, 281  
Liakata, Maria, 539  
Lima, Rogério Braunschweiger de Freitas, 441  
Lin, Ying-Jia, 362  
Lindevelt, David, 82  
Liu, Hongfang, 551  
Livingstone, Elisabeth, 455  
Loane, Russell, 296  
Lovon-Melgarejo, Jesus, 384  
Ludmir, Ethan, 139

Magnini, Bernardo, 245  
Majeedi, Abrar, 530  
Maniatis, Alexandros, 441  
Margolina, Anastasia, 41  
Marimuthu, Sushvin, 268  
Martín-Valdivia, María-Teresa, 72  
Martinc, Matej, 223  
Masud, Md Akmol, 506  
Matos, Sérgio, 441  
Matthes, Florian, 402, 469  
Mishra, Sumit, 390  
Molino-Piñar, Lucas, 72  
Mondal, Anindita, 395  
Mulcahey, Rory, 296

Nagai, Hiroyuki, 107  
Nensa, Felix, 455  
Ng, Yee Man, 15

Nishiyama, Tomohiro, 107, 149  
Ntanavaras, Stergios, 171

Ojeda Trueba, Sergio LUIS, 159  
Oronoz, Maite, 434

Pakull, Tabea Margareta Grace, 455  
Pantin, Jérémie, 384  
Parde, Natalie, 487  
Pérez, Alicia, 434  
Pham, Luan Huy, 281  
Pitarch, Lucia, 159  
Puts, Sander, 82

Radoi, Emilian, 132  
Rai, Siddhant, 530  
Rajwal, Swati, 307  
Ranjan, Ravi, 428  
Rao, Samminga Sainath, 390  
Reynolds, Tera L, 506  
Roberts, Kirk, 139  
Rusnachenko, Nicolay, 482

Salvi, Rohan Charudatt, 320  
Sánchez-Montero, Alec, 159  
Sansaloni-Pastor, Sara, 234  
Santhosh, Roshan, 428  
Sarker, Abeed, 307  
Sharma, Dipti Misra, 268  
Simon, Betty, 268  
Sola-Garcia, Elena, 72  
Sommer, Katharina, 402  
Soni, Sarvesh, 255  
Soufleri, Efstathia, 190  
Spruit, Marco, 15  
Stöckl, Andreas, 96, 117  
Su, Xinqi, 551  
Suman, Chanchal, 390  
Sun, Mengxuan, 482

Tänzer, Michael, 26  
Telnova, Alina, 41  
Teodoro, Douglas, 234  
Thompson, Paul, 190  
Till, Tristan, 402  
Torri, Vittorio, 412  
Tran, Tai Tan, 497  
Tripp, Bryan, 1  
Tseriotou, Talia, 539

Ulli, Vinay Babu, 395  
Urbano-Moral, Jose Angel, 72  
Uzuner, Özlem, 281

van Dijk, Bram, 15  
van Duijn, Max J., 15  
van Mulligen, Erik, 82  
van Oortmerssen, Gerard, 15  
Vedde, Sol Gedde, 61  
Veloso, João Pedro, 50  
Verberne, Suzan, 82  
Vossen, Piek T.J.M., 171

Wakamiya, Shoko, 107, 149  
Wang, Rongrong, 551  
Wihardjo, Inez, 565

Yadav, Shweta, 320  
Yamada, Keiko, 107  
Yang, Yumeng, 139  
Yazdani, Anthony, 234  
Yurt, Ibrahim Ebrar, 469

Zabolotnii, Serhii, 375  
Zemchyk, Arina, 524