LREC 2026

The 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026

Workshop Proceedings

Editors
Asma Ben Abacha
Steven Bethard
Danielle Bitterman
Tristan Naumann
Kirk Roberts

16 May 2026

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 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 978-2-493814-56-2

Preface

This volume contains papers from the 8th Workshop on Clinical Natural Language Processing (Clinical NLP 2026).

Clinical text is growing rapidly as electronic health records (EHRs) become pervasive. Much of the information recorded in a clinical encounter is located exclusively in provider narrative notes, which makes them indispensable for supplementing structured clinical data to better understand patient state and care provided. The goal of this workshop is to establish a regular event that brings together researchers interested in developing state-of-the-art methods for the clinical domain. The focus is on improving NLP technology to enable clinical applications via information extraction and modeling of narrative provider notes from electronic health records, patient encounter transcripts, and other clinical narratives. Clinical NLP 2026 also hosted two shared tasks, challenging researchers around the world to develop new approaches to solve clinical NLP problems: scoring model-generated answers to patient questions and extracting/normalizing clinical observations from conversational transcripts.

The main workshop received a total of 40 submissions, of which 7 were accepted as oral presentations, and 14 were accepted as poster presentations. The shared tasks received a total of 18 participant submissions, of which 2 were accepted as oral presentations, and 16 were accepted as poster presentations.

Keynote Talk

Holistic engagement beyond the edges of data-driven science: The case of Libya

Dr. Stephen Wu
Saraya Hamra University

Abstract: We often make some assumptions about the ethics (e.g., individual rights) and structures (e.g., research enterprise) supporting health technologies. This talk argues that, for societies like Libya's, some paradigm shifts are in order. First, we contrast ethical values between honor vs. dignity cultures. Second, we highlight aspects of the Libyan research system that fundamentally change what work needs to be done in health technology research. Subsequently, we describe several recent holistic, contextually situated efforts at engagement: private research sponsorship and administration, science of science research, and care practice norms and information flow in medicine.

Biography: Dr. Stephen Wu is an Associate Professor at Saraya Hamra University (SHU) in Tripoli, Libya, where he serves as the Director for the Center for Research & Innovation as well as the Head of the Department of Computer Science & Engineering. His current interests lie in empowering the Libyan innovation system through the science of science, the interaction of global value systems with modern AI, and culturally situated healthcare information systems.

Organizing Committee

General Chairs

Asma Ben Abacha, Microsoft, USA

Steven Bethard, University of Arizona, USA

Danielle Bitterman, Harvard University, USA

Tristan Naumann, Microsoft Research, USA

Kirk Roberts, University of Texas Health Science Center at Houston, USA

MEDIQA-EVAL 2026 Shared Task Organizers

Asma Ben Abacha, Microsoft, USA

Wen-wai Yim, Microsoft, USA

MEDIQA-SYNUR 2026 Shared Task Organizers

Asma Ben Abacha, Microsoft, USA

Nate Bodenstab, Microsoft, USA

Jean-Philippe Corbeil, Microsoft, Canada

George Michalopoulos, Microsoft, Canada

Reviewers

Abdulrahman AAlAbdulsalam, Bayu Aryoyudanta, Federica Bologna, Leonardo Campillos-Llanos, Jean-Philippe Corbeil, Yadan Fan, Haritha Gangavarapu, Natalia Grabar, Ziqi Hao, Harry Hochheiser, Sy Hwang, Kyomin Hwang, Yoshinobu Kano, Halil Kilicoglu, Primoz Kocbek, Yanis Labrak, Egoitz Laparra, Alberto Lavelli, Ulf Leser, Yuxing Lu, Diwakar Mahajan, Sushvin Marimuthu, George Michalopoulos, Timothy Miller, Prateek Munjal, Aurélie Névéol, Yifan Peng, Xinzhe Peng, Lisa Raithel, Satyajeet Raje, Pavithra Rajendran, Giridhar Kaushik Ramachandran, A H M Rezaul Karim, Frank Rudzicz, Michael Saban, Guergana Savova, Ankit Singh, Arvind Krishna Sridhar, Swetha Krishna Sriram, Lynda Tamine, Jerrin Thomas, Aashwin Vaish, Airat Valiev, V.G.Vinod Vydiswaran, Jinge Wu, Justin Xu, Dongfang Xu, Jiarui Yao, Wen-wai Yim, WonJin Yoon, Boya Zhang, Pierre Zweigenbaum

Table of Contents

Overview of the MEDIQA-EVAL 2026 Shared Task on Evaluation Metrics in Medical Multimodal Question Answering
Asma Ben Abacha and Wen-wai Yim 1

SUAT-BMI at MEDIQA-EVAL 2026: An Ensemble Approach to Language Models as Judges for Automatic Rating of Medical Responses
Xinzhe Peng, Liyuan E, Kun Feng, Jielin Li, Yuxuan Tang and Zhao Li 12

Overview of the MEDIQA-SYNUR 2026 Shared Task on Observation Extraction from Nurse Dictations
George Michalopoulos, Jean-Philippe Corbeil, Cari Bader, Nathan Bodenstab and Asma Ben Abacha 19

SemAnTICA Lab at MediQA-SYNUR 2026: Route, Extract and Verify – An LLM-gated Ensemble for Parsing Nurse Dictations
Sy Hwang, Katherine S. Pitcher, Sue Hyon Kim, Yoonjae Lee, Hayoung K. Donelly, Harsh Bandhey, Andrew J. King, Karen O'Connor, Ryan J. Urbanowicz and Danielle L. Mowery . . . 27

L2D-Clinical: Learning to Defer for Adaptive Model Selection in Clinical Text Classification
Rishik Kondadadi and John E. Ortega 38

TRUMEDIQA: A Modular Trustworthy RAG Pipeline for Multilingual Medical Question Answering
Meryem El Fatimi, Ayoub Nainia and Jihad Zahir 49

Evaluating the Retrieval Component in a Retrieval-Augmented Summarization System for Patient Records in French
Marco Naguib, Christel Gérardin, Victor Beaucoté, Cyril Charron, Adrien Joseph, Aurélie Névéol and Xavier Tannier 57

Retrieval-Augmented Generation Based Nurse Observation Extraction
Kyomin Hwang and Nojun Kwak 66

Automatic Generation of Discharge Summaries Using Large Language Models: A Systematic Literature Review
Lucas Molino-Piñar, Manuel Carlos Diaz Galiano and María-Teresa Martín-Valdivia 73

Smart_solutions at MEDIQA-SYNUR 2026: A Multi-Stage LLM Pipeline for Nursing Observation Extraction
Prateek Munjal 83

GS-BrainText: A Multi-Site Brain Imaging Report Dataset from Generation Scotland for Clinical Natural Language Processing Development and Validation
Beatrice Alex, Claire Grover, Arlene Casey, Richard Tobin, Heather Whalley and William Whiteley 92

SASTA Self Assessment: An efficient human-in-the-loop strategy for developmental and pathological language analysis
Jan Odijk, Jelte van Boheemen, Xander Vertegaal, Tessel Boerma and Marijn Schraagen 103

Differentially Private De-identification of Dutch Clinical Notes: A Comparative Evaluation
Michele Miranda, Xinlan Yan, Nishant Mishra, Rachel Murphy, Ameen Abu Hanna, Sébastien Bratières and Iacer Calixto 113

SQUCS at MEDIQA-SYNUR 2026: A Multi-Agent Open Source LLM System for Nursing Observation Extraction
Riham JeebAllah, Adhari AlZaabi and Abdulrahman Khalifa AAlAbdulsalam 125

LTRC-IIIT at MEDIQA-SYNUR 2026: Benchmarking a Fully Local, Training-Free RAG Pipeline
Aashwin Vaish and Dipti Misra Sharma 136

BDI at MEDIQA-EVAL 2026: A ReAct-Style Multimodal Agent for Fine-Grained Medical Response Assessment
Justin Xu, Zizheng Zhang, Augustine Luk, Benjamin Khong, Haochen Cui, Samuel Hwang, Alyssa Pradhan, Kevin Yuan and David W. Eyre 141

MasonNLP at MEDIQA-SYNUR 2026: Retrieval-Augmented Large Language Models for Schema-Constrained Clinical Information Extraction
A H M Rezaul Karim and Özlem Uzuner 153

Night Shift Nerds at MEDIQA-SYNUR 2026: Pushing Small Large Language Model Capability for Clinical Observation Extraction and Normalization from Nurse Dictation using RLVR
Bayu Aryoyudanta, Maria Yuliana, Mikie Rachman and I Made Agus Setiawan 163

Extracting Medication Instructions from Dutch General Practice Electronic Health Records with Local Natural Language Processing
Marya Dukmak, Constanza L. Andaur Navarro and Artuur Leeuwenberg 174

Gladiator at MEDIQA-SYNUR 2026: Contextual Clinical Extraction: Integrating Foundation Models with Domain-Specific Validation Rules
Siva Satyanarayana Raju Pusapati and Ankit Singh 183

MedAware at MEDIQA-EVAL 2026: Vision-Language Model Fine-Tuning with Logprob-Based Score Calibration for Medical Response Evaluation
Ziqi Hao and Pengbo Liu 192

HSE NLP TEAM at MEDIQA-SYNUR 2026: Consensus Adjudication Ensemble (ACE): Balancing Precision and Recall for Schema-Bystander Clinical Extraction
Airat A. Valiev 200

Lakefront AI Ramblers at MEDIQA-SYNUR 2026: Hybrid Retrieval and LLM Verification for Open-Source Schema-Guided Clinical Information Extraction
Michael T. Saban, Arsalan Yaghoubi, Behnaz Eslami, Samie Tootooni and Dmitriy Dligach 212

JMedWiC: A Japanese Word-in-Context Dataset in the Medical Domain
Koki Horiguchi, Seiji Sugiyama, Tomoyuki Kajiwara, Shoko Wakamiya and Eiji Aramaki222

LTRC-Medicom at MEDIQA-SYNUR 2026: Schema-Guided Clinical Information Extraction with Hybrid Clustering-SFT-Verification
Pasumarthy Deepak, Sushvin Marimuthu and Parameswari Krishnamurthy 228

SloCal-Net at MEDIQA-Eval 2026: Investigating the Impact of Reasoning and External Context on Medical Answer Grading
Primoz Kocbek, Valentina Carbonari, Pierangelo Veltri, Pietro Hiram Guzzi and Gregor Stiglic 235

MIDAS_SYNUR at MEDIQA-SYNUR 2026: A Prompting Study for Clinical Observation Extraction from Nurse Dictation Transcriptions
Swetha Krishna Sriram and Akshitaa Sahoo 244

AnotherOne at MEDIQA-SYNUR 2026: Detect, Extract, Normalize - Knowledge-Grounded LLM Pipeline for Clinical Observation Extraction
Jerrin John Thomas and Parameswari Krishnamurthy 251

hgkai26 at MEDIQA-EVAL 2026: Automated Evaluation of Visual Medical Question Answering Using LLM-as-a-Judge
Haritha Gangavarapu 257

Role-Adapted Clinical Report Generation for Ultrasound Measurements in Low-Resource Settings
Ayoub Nainia, Tanya Akumu, Noussair Lazrak and Karim Lekadir 262

A Comparative Study of Approaches to Anonymization of Clinical Free Text in Spanish
Florencia Luciana Brunello, Laura Alonso Alemany, Serena Villata and Milagro Teruel 270

Disagreement-Driven Joint Refinement of Retrieval and Decision Rules for Imbalanced Counseling Risk Classification
Zhihao Shao, Ryo Sekizaki, Shengzhou Yi and Toshihiko Yamasaki 281

Context-Aware SNOMED CT Entity Linking for Clinical Text
Provia Kadusabe, Demian Gholipour Ghalandari, Lauren Cassidy, Jack Boylan, Chris Hokamp, Abhishek Kaushik and Fiona Lawless 290

Temporal Structure in Clinical Narratives in Portuguese: Insights from Cross-Document Annotation
Ana Luisa Fernandes, Purificação Silvano and Luís Filipe Cunha 300

MOSAIC: A Multilingual, Taxonomy-Agnostic, and Computationally Efficient Approach for Radiological Report Classification in Low-Resource Settings
Alice Schiavone, Marco Fraccaro, Lea Marie Pehrson, Silvia Ingala, Rasmus Bonnevie, Michael Bachmann Nielsen, Vincent Beliveau, Melanie Ganz and Desmond Elliott 313

MedNormJ: A Benchmark Dataset for Medical Concept Normalization in Japanese Clinical Documents
Yuki Tashiro, Seiji Shimizu, Tomohiro Nishiyama, Shoko Wakamiya and Eiji Aramaki . 324

Pediatric Sepsis Cohort Detection Using In-Context Pointwise V-Usable Information
Yingya Li, Alon Geva, Steven Bethard, Timothy A. Miller, Kate Madden, Matthew A. Eisenberg, Daniel P. Kelly and Guergana Savova 336

RECAP: Transparent Inference-Time Emotion Alignment for Medical Dialogue Systems
Adarsh Srinivasan, Jacob Dineen, Muhammad Uzair Sarfraz, Muhammad Umar Afzal, Irbaz Riaz and Ben Zhou 350

Disentangling Ambiguity from Instability in Large Language Models: A Clinical Text-to-SQL Case Study
Angelo Ziletti and Leonardo D'Ambrosi 369

An OMOP-Based Open-Source Text-to-SQL Benchmark Dataset
Paul Legrand, Kawsar Noor, Satyam Bhagwanani and Richard J. Dobson 381

Profiling Hallucinations in Frontier LLMs for Entity Linking to Medical Ontologies
Logan Born, Nishant Kambhatla, Uliyana Kubasova, Maryam Siahbani, Andrei Vacariu, Timothy W. O'Connell and Anoop Sarkar 394

Conference Program

Sat 16 May 2026

09:00–09:45 **Opening**

09:00–09:30 *Keynote: Holistic engagement beyond the edges of data-driven science: The case of Libya*
Dr. Stephen Wu

09:30–09:45 ***Keynote Q&A***

09:45–10:00 **MEDIQA-EVAL 2026 Shared Task**

09:45–09:55 *Overview of the MEDIQA-EVAL 2026 Shared Task on Evaluation Metrics in Medical Multimodal Question Answering*
Asma Ben Abacha and Wen-wai Yim

09:55–10:00 *SUAT-BMI at MEDIQA-EVAL 2026: An Ensemble Approach to Language Models as Judges for Automatic Rating of Medical Responses*
Xinzhe Peng, Liyuan E, Kun Feng, Jielin Li, Yuxuan Tang and Zhao Li

10:00–10:15 **MEDIQA-SYNUR 2026 Shared Task**

10:00–10:10 *Overview of the MEDIQA-SYNUR 2026 Shared Task on Observation Extraction from Nurse Dictations*
George Michalopoulos, Jean-Philippe Corbeil, Cari Bader, Nathan Bodenstab and Asma Ben Abacha

10:10–10:15 *SemAnTICA Lab at MediQA-SYNUR 2026: Route, Extract and Verify – An LLM-gated Ensemble for Parsing Nurse Dictations*
Sy Hwang, Katherine S. Pitcher, Sue Hyon Kim, Yoonjae Lee, Hayoung K. Donelly, Harsh Bandhey, Andrew J. King, Karen O'Connor, Ryan J. Urbanowicz and Danielle L. Mowery

10:15–11:15 **Poster Session**

10:15–11:15 *L2D-Clinical: Learning to Defer for Adaptive Model Selection in Clinical Text Classification*
Rishik Kondadadi and John E. Ortega

10:15–11:15 *TRUMEDIQA: A Modular Trustworthy RAG Pipeline for Multilingual Medical Question Answering*
Meryem El Fatimi, Ayoub Nainia and Jihad Zahir

Sat 16 May 2026 (continued)

10:15–11:15 *Evaluating the Retrieval Component in a Retrieval-Augmented Summarization System for Patient Records in French*
Marco Naguib, Christel Gérardin, Victor Beaucoté, Cyril Charron, Adrien Joseph, Aurélie Névéol and Xavier Tannier

10:15–11:15 *Retrieval-Augmented Generation Based Nurse Observation Extraction*
Kyomin Hwang and Nojun Kwak

10:15–11:15 *Automatic Generation of Discharge Summaries Using Large Language Models: A Systematic Literature Review*
Lucas Molino-Piñar, Manuel Carlos Diaz Galiano and María-Teresa Martín-Valdivia

10:15–11:15 *Smart_solutions at MEDIQA-SYNUR 2026: A Multi-Stage LLM Pipeline for Nursing Observation Extraction*
Prateek Munjal

10:15–11:15 *GS-BrainText: A Multi-Site Brain Imaging Report Dataset from Generation Scotland for Clinical Natural Language Processing Development and Validation*
Beatrice Alex, Claire Grover, Arlene Casey, Richard Tobin, Heather Whalley and William Whiteley

10:15–11:15 *SASTA Self Assessment: An efficient human-in-the-loop strategy for developmental and pathological language analysis*
Jan Odijk, Jelte van Boheemen, Xander Vertegaal, Tessel Boerma and Marijn Schraagen

10:15–11:15 *Differentially Private De-identification of Dutch Clinical Notes: A Comparative Evaluation*
Michele Miranda, Xinlan Yan, Nishant Mishra, Rachel Murphy, Ameen Abu Hanna, Sébastien Bratières and Iacer Calixto

10:15–11:15 *SQUCS at MEDIQA-SYNUR 2026: A Multi-Agent Open Source LLM System for Nursing Observation Extraction*
Riham JeebAllah, Adhari AlZaabi and Abdulrahman Khalifa AAlAbdulsalam

10:15–11:15 *LTRC-IIIT at MEDIQA-SYNUR 2026: Benchmarking a Fully Local, Training-Free RAG Pipeline*
Aashwin Vaish and Dipti Misra Sharma

10:15–11:15 *BDI at MEDIQA-EVAL 2026: A ReAct-Style Multimodal Agent for Fine-Grained Medical Response Assessment*
Justin Xu, Zizheng Zhang, Augustine Luk, Benjamin Khong, Haochen Cui, Samuel Hwang, Alyssa Pradhan, Kevin Yuan and David W. Eyre

Sat 16 May 2026 (continued)

10:15–11:15 *MasonNLP at MEDIQA-SYNUR 2026: Retrieval-Augmented Large Language Models for Schema-Constrained Clinical Information Extraction*
A H M Rezaul Karim and Özlem Uzuner

10:15–11:15 *Night Shift Nerds at MEDIQA-SYNUR 2026: Pushing Small Large Language Model Capability for Clinical Observation Extraction and Normalization from Nurse Dictation using RLVR*
Bayu Aryoyudanta, Maria Yuliana, Mikie Rachman and I Made Agus Setiawan

10:15–11:15 *Extracting Medication Instructions from Dutch General Practice Electronic Health Records with Local Natural Language Processing*
Marya Dukmak, Constanza L. Andaur Navarro and Artuur Leeuwenberg

10:15–11:15 *Gladiator at MEDIQA-SYNUR 2026: Contextual Clinical Extraction: Integrating Foundation Models with Domain-Specific Validation Rules*
Siva Satyanarayana Raju Pusapati and Ankit Singh

10:15–11:15 *MedAware at MEDIQA-EVAL 2026: Vision-Language Model Fine-Tuning with Logprob-Based Score Calibration for Medical Response Evaluation*
Ziqi Hao and Pengbo Liu

10:15–11:15 *HSE NLP TEAM at MEDIQA-SYNUR 2026: Consensus Adjudication Ensemble (ACE): Balancing Precision and Recall for Schema-Bystander Clinical Extraction*
Airat A. Valiev

10:15–11:15 *Lakefront AI Ramblers at MEDIQA-SYNUR 2026: Hybrid Retrieval and LLM Verification for Open-Source Schema-Guided Clinical Information Extraction*
Michael T. Saban, Arsalan Yaghoubi, Behnaz Eslami, Samie Tootooni and Dmitriy Dligach

10:15–11:15 *JMedWiC: A Japanese Word-in-Context Dataset in the Medical Domain*
Koki Horiguchi, Seiji Sugiyama, Tomoyuki Kajiwara, Shoko Wakamiya and Eiji Aramaki

10:15–11:15 *LTRC-Medicom at MEDIQA-SYNUR 2026: Schema-Guided Clinical Information Extraction with Hybrid Clustering-SFT-Verification*
Pasumarthy Deepak, Sushvin Marimuthu and Parameswari Krishnamurthy

10:15–11:15 *SloCal-Net at MEDIQA-Eval 2026: Investigating the Impact of Reasoning and External Context on Medical Answer Grading*
Primoz Kocbek, Valentina Carbonari, Pierangelo Veltri, Pietro Hiram Guzzi and Gregor Stiglic

10:15–11:15 *MIDAS_SYNUR at MEDIQA-SYNUR 2026: A Prompting Study for Clinical Observation Extraction from Nurse Dictation Transcriptions*
Swetha Krishna Sriram and Akshitaa Sahoo

Sat 16 May 2026 (continued)

10:15–11:15 *AnotherOne at MEDIQA-SYNUR 2026: Detect, Extract, Normalize - Knowledge-Grounded LLM Pipeline for Clinical Observation Extraction*
Jerrin John Thomas and Parameswari Krishnamurthy

10:15–11:15 *hgkai26 at MEDIQA-EVAL 2026: Automated Evaluation of Visual Medical Question Answering Using LLM-as-a-Judge*
Haritha Gangavarapu

10:15–11:15 *Role-Adapted Clinical Report Generation for Ultrasound Measurements in Low-Resource Settings*
Ayoub Nainia, Tanya Akumu, Noussair Lazrak and Karim Lekadir

10:15–11:15 *A Comparative Study of Approaches to Anonymization of Clinical Free Text in Spanish*
Florencia Luciana Brunello, Laura Alonso Alemany, Serena Villata and Milagro Teruel

10:15–11:15 *Disagreement-Driven Joint Refinement of Retrieval and Decision Rules for Imbalanced Counseling Risk Classification*
Zhihao Shao, Ryo Sekizaki, Shengzhou Yi and Toshihiko Yamasaki

10:15–11:15 *Context-Aware SNOMED CT Entity Linking for Clinical Text*
Provia Kadusabe, Demian Gholipour Ghalandari, Lauren Cassidy, Jack Boylan, Chris Hokamp, Abhishek Kaushik and Fiona Lawless

10:15–11:15 *Temporal Structure in Clinical Narratives in Portuguese: Insights from Cross-Document Annotation*
Ana Luisa Fernandes, Purificação Silvano and Luís Filipe Cunha

11:15–11:45 Special Track on Low Resource Settings

11:15–11:30 *MOSAIC: A Multilingual, Taxonomy-Agnostic, and Computationally Efficient Approach for Radiological Report Classification in Low-Resource Settings*
Alice Schiavone, Marco Fraccaro, Lea Marie Pehrson, Silvia Ingala, Rasmus Bonnevie, Michael Bachmann Nielsen, Vincent Beliveau, Melanie Ganz and Desmond Elliott

11:30–11:45 *MedNormJ: A Benchmark Dataset for Medical Concept Normalization in Japanese Clinical Documents*
Yuki Tashiro, Seiji Shimizu, Tomohiro Nishiyama, Shoko Wakamiya and Eiji Aramaki

Sat 16 May 2026 (continued)

11:45–12:15 **Special Track on Construction or Deconstruction of LLMs**

11:45–12:00 *Pediatric Sepsis Cohort Detection Using In-Context Pointwise V-Usable Information*
Yingya Li, Alon Geva, Steven Bethard, Timothy A. Miller, Kate Madden, Matthew A. Eisenberg, Daniel P. Kelly and Guergana Savova

12:00–12:15 *RECAP: Transparent Inference-Time Emotion Alignment for Medical Dialogue Systems*
Adarsh Srinivasan, Jacob Dineen, Muhammad Uzair Sarfraz, Muhammad Umar Afzal, Irbaz Riaz and Ben Zhou

12:15–13:00 **Main Track**

12:15–12:30 *Disentangling Ambiguity from Instability in Large Language Models: A Clinical Text-to-SQL Case Study*
Angelo Ziletti and Leonardo D'Ambrosi

12:30–12:45 *An OMOP-Based Open-Source Text-to-SQL Benchmark Dataset*
Paul Legrand, Kawsar Noor, Satyam Bhagwanani and Richard J. Dobson

12:45–13:00 *Profiling Hallucinations in Frontier LLMs for Entity Linking to Medical Ontologies*
Logan Born, Nishant Kambhatla, Uliyana Kubasova, Maryam Siahbani, Andrei Vacariu, Timothy W. O'Connell and Anoop Sarkar

13:00 **Closing**

Overview of the MEDIQA-EVAL 2026 Shared Task on Evaluation Metrics in Medical Multimodal Question Answering

Asma Ben Abacha, Wen-wai Yim
{abenabacha, yimwenwai}@microsoft.com
Microsoft, Health and Life Sciences AI, Redmond, US

Abstract

Evaluating clinical text generation remains challenging, as automatic metrics often correlate weakly with clinician judgments. This issue is particularly pronounced in medical multimodal question answering (MMQA), where systems must integrate visual and textual information and evaluation must capture factual accuracy, visual grounding, completeness, and overall coherence. Despite rapid progress in MMQA, there is limited consensus on clinically meaningful evaluation, and existing metrics, largely adapted from general NLG or VQA, often fail to capture domain-specific criteria. We introduce MEDIQA-EVAL 2026, a shared task on evaluation metrics for medical multimodal QA. To our knowledge, this is the first shared task focused on evaluating automatic metrics in this setting. We release a dataset of medical visual question-answer pairs annotated with multidimensional clinician judgments. Systems are evaluated by the correlation of their metric scores with expert ratings on a held-out test set. Participants explored diverse approaches, including vision-language models, retrieval-augmented judging, metric-specific classifiers, reinforcement learning, and LLM-as-a-judge frameworks. Results show that model-based evaluators achieve stronger alignment with human judgments than traditional NLG metrics, particularly on English data, while performance remains lower on Chinese, highlighting challenges in multilingual evaluation. Notably, our MEDIQA LLM-as-a-judge approach achieves strong performance across both languages.

Keywords: Evaluation, Medical Multimodal Question Answering, LLM-as-a-Judge, Human Alignment

1. Introduction

Evaluating clinical text generation remains challenging, particularly in medical settings where assessments may vary across experts and correctness is often graded rather than strictly binary. Widely used reference-based automatic metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and BLEURT (Sellam et al., 2020) have been shown to correlate weakly with clinician judgments in clinical contexts. Studies on automated medical note generation further demonstrates that standard metrics do not reliably capture clinically important properties such as factual accuracy, omissions, and hallucinations, and that domain-adapted or composite metrics may better approximate expert assessment (Ben Abacha et al., 2023).

LLM-based evaluators such as GPTScore (Fu et al., 2024) have been proposed to assess generated text along quality dimensions, offering greater flexibility for open-ended generation tasks. More recently, the MORQA benchmark for medical open-response QA finds that LLM-as-a-judge approaches (e.g., GPT-4 and Gemini) tend to correlate more strongly with expert ratings than traditional metrics across multilingual, multimodal QA datasets, in part because they are more sensitive to semantic nuances and robust to variability among multiple valid references (Yim et al., 2026). These insights highlight the need for evaluation methods that go beyond surface overlap and better align with clinical judgment in complex QA scenarios.

The challenge is further amplified by advances in Vision–Language Models (VLMs). Although VLMs can generate fluent, human-like responses, reliably evaluating the accuracy and completeness of their outputs remains difficult, particularly in multimodal settings where images must be interpreted alongside textual information and where multiple clinically acceptable responses may exist. In medical multimodal question answering (MMQA), systems must integrate visual evidence and clinical language, yet evaluation practices largely remain adapted from general natural language generation and VQA. While datasets such as VQA-RAD (Lau et al., 2018), VQA-Med (Ben Abacha et al., 2019, 2020, 2021), and PathVQA (He et al., 2020) have enabled rapid model development, evaluation has often relied on exact-match or classification-style accuracy for short, constrained answers. Such setups only partially reflect clinical needs in open-ended MMQA.

To address this gap, we introduce the MEDIQA-EVAL 2026 shared task on Evaluation Metrics for Medical Multimodal Question Answering. We extend the MEDIQA-M3G 2024, MEDIQA-MAGIC 2024, MEDIQA-WV 2025, and MEDIQA-MAGIC-2025 (Yim et al., 2024a,b, 2025c,a) shared tasks on dermatology and wound-care visual question answering by shifting the focus to evaluating open-ended system responses. In this multimodal evaluation task, participants develop automatic metrics that assign quality scores to model-generated answers for patient questions paired with one or multiple clinical images. Responses are assessed

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 1–11
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Figure 1: Overview of the meta-evaluation pipeline used in the MEDIQA-EVAL 2026 shared task. Stage 1 provides the multimodal dataset with patients' questions and medical answers, system-generated answers, and clinician reference answers. In Stage 2, participants submit automatic metric scores computed over the provided system-generated answers. Stage 3 evaluates metric validity by measuring correlation between metric scores and clinician quality ratings. Stage 4 ranks submitted metrics on a leaderboard based on their alignment with clinician judgments.

across multiple clinically motivated quality dimensions, including completeness, factual accuracy, relevance, writing style, and overall quality.

We provide clinician ratings for the development set and rank submitted metrics based on their correlation with expert judgments on the test set. By centering alignment with clinician assessments and adopting a multidimensional evaluation framework, this shared task establishes a benchmark for MMQA metric evaluation and promotes the development of reliable, clinically grounded evaluation methods for multimodal clinical QA systems.

2. Task Description

We introduce a clinician-aligned meta-evaluation framework for assessing automatic evaluation metrics in medical question answering. Unlike conventional shared tasks that rank model outputs, the MEDIQA-EVAL shared task[1] evaluates the quality of evaluation metrics themselves. The primary objective is to determine the extent to which automatic metrics align with expert clinician judgments. Figure 1 illustrates the overall evaluation pipeline. Rather than submitting system predictions, participants submit metric-generated scores, which are subsequently evaluated based on their agreement with clinician ratings.

2.1. Stage 1 - Task Data

The MEDIQA-EVAL 2026 dataset[2] spans two languages (English and Chinese) and consists of:

- Patient questions and associated medical images,
- System-generated answers produced by multimodal medical QA systems, and associated clinician quality ratings (provided in the development set only),
- Clinician reference answers.

Clinician ratings serve as the gold standard for metric validation. These annotations are multidimensional and language-specific. For the English datasets, responses are evaluated across several quality dimensions, including completeness, factual accuracy, relevance, writing style, and overall quality. For the Chinese datasets, domain experts assess factual consistency with the gold standard and writing style.

2.2. Stage 2 - Response Scoring

Participants are provided with the dataset described above and are required to compute quality scores using their proposed evaluation metric. For each system-generated answer, their proposed metric must output a score reflecting the estimated response quality.

Participants do not submit model outputs. Instead, they submit metric-generated scores computed over the provided system answers to ensure that all methods are evaluated on the same inputs.

2.3. Stage 3 - Clinician Alignment

To assess metric quality, we measure the statistical correlation between participant-submitted metric scores and clinician quality ratings (Figure 1, Stage 3). Because clinician evaluation spans multiple quality dimensions, we compute correlations both at the dimension level and for overall quality:

- For the EN datasets, clinician ratings include the following dimensions: *disagreement flag*, *completeness*, *factual accuracy*, *relevance*, *writing style*, and *overall quality*.
- For the ZH datasets, ratings include *factual consistency with the gold standard* and *writing style*.

[1] https://sites.google.com/view/mediqa2026/mediqa-eval

[2] The MEDIQA-EVAL 2026 dataset is available at https://osf.io/kcv2n

Let m_i denote the score assigned by a metric to instance i, and let $c_i^{(d)}$ denote the clinician rating for instance i on dimension d (e.g., completeness, factual accuracy, writing style, or overall quality). Metric validity with respect to a given dimension d is quantified as:

$$\text{Alignment}_d(m) = \text{corr}(m_i, c_i^{(d)}) \quad (1)$$

In addition, we report alignment with the overall clinician rating when available. Correlation is computed using **Pearson's** r, **Spearman's** ρ, and **Kendall's** τ. Higher correlation indicates stronger agreement with clinician judgment on the evaluated dimension and, therefore, greater metric reliability.

2.4. Stage 4 - Leaderboard & Results

Metrics are ranked according to their correlation scores (Figure 1, Stage 4). Specifically, leaderboard positions are determined by the strength of alignment between metric scores and expert annotations across the evaluated quality dimensions. The competition therefore identifies evaluation methods that most faithfully approximate expert clinical judgment. By grounding metric validation in multidimensional clinician assessment, MEDIQA-EVAL moves beyond surface-level similarity measures and prioritizes evaluation approaches that better capture clinical relevance, factual accuracy, completeness, and safety-related considerations.

3. Datasets & Annotations

We use the MORQA benchmark (Yim et al., 2026), which is built upon two dermatology-focused visual question answering (VQA) datasets: WoundcareVQA (Yim et al., 2025b) and DermaVQA-iiyi (Yim et al., 2024c). Both source datasets contain English and Chinese question-answer pairs and are centered on dermatology, a visually grounded medical specialty in which clinical assessment heavily depends on images. Each instance consists of a patient question paired with corresponding medical image(s) and clinician-provided reference answers.

In the MORQA benchmark[3], we further augment this data with system-generated answers from multimodal medical QA models and associated expert quality ratings. These additional annotations enable direct evaluation of automatic metrics by measuring their alignment with clinician judgments.

3.1. English Datasets

For the English (WoundcareVQA and DermaVQA-iiyi) datasets, each system-generated response was evaluated by a practicing medical doctor using a multidimensional annotation framework. Annotations include a binary disagreement flag (`disagree_flag`) indicating whether the expert disagreed with the response, as well as graded assessments of `completeness`, `factual_accuracy`, `relevance`, `writing_style`, and `overall`.

All graded dimensions are scored on a three-point scale $\{0, 0.5, 1.0\}$, where 1.0 denotes a fully satisfactory response, 0.5 denotes a partially satisfactory response, and 0 denotes an inaccurate, incomplete, or otherwise inadequate response.

3.2. Chinese Datasets

For the Chinese (WoundcareVQA and DermaVQA-iiyi) datasets, each system-generated response was evaluated by a single domain expert trained at a Chinese medical school. The annotation framework includes two dimensions: `factual_consistency_wgold` and `writing_style`.

`factual_consistency_wgold` measures the degree to which a system response is factually consistent with the clinician-provided reference (gold) answer, while `writing_style` assesses the appropriateness and clarity of the response. Both dimensions are scored on a three-point scale $\{0, 0.5, 1.0\}$, where 1.0 indicates full satisfaction, 0.5 indicates partial satisfaction, and 0 indicates an incorrect, inconsistent, or inappropriate response.

3.3. Evaluation Metrics (English)

Following Section 2.3 (Stage 3 - Clinician Alignment), all submitted metrics are evaluated by measuring their correlation with clinician ratings under the evaluation script provided by the task[4]. Correlation is computed using Kendall's τ, Pearson's r, and Spearman's ρ.

For the English datasets, alignment is computed separately for each annotated quality dimension, reflecting the multidimensional clinician evaluation framework. In total, 73 metrics are reported, defined by:

- **Dataset**: {iiyi, woundcare, ALL}
- **Evaluation Dimension (EN_METRIC)**: {en-disagree_flag, en-completeness, en-factual-accuracy, en-relevance, en-writing-style, en-overall}

For each Dataset–EN_METRIC pair, the following correlation scores are computed:

1. Dataset-EN_METRIC-kendalltau

[3] https://osf.io/kcv2n

[4] https://github.com/wyim/MEDIQA-EVAL-2026

2. Dataset-EN_METRIC-pearson
3. Dataset-EN_METRIC-spearman
4. Dataset-EN_METRIC-mean

where `mean` denotes the average of the three correlation coefficients. In addition, we report an aggregated score (ALL-en-ALL-mean), computed by averaging correlations across all English datasets and evaluation dimensions.

3.4. Evaluation Metrics (Chinese)

For the Chinese datasets, metric validity is likewise assessed via correlation with clinician annotations, as described in Stage 3. Given the language-specific annotation framework, alignment is computed with respect to two quality dimensions. In total, 25 metrics are reported, defined by:

- **Dataset**: {iiyi, woundcare, ALL}
- **Evaluation Dimension (ZH_METRIC)**: {zh-factual-consistency-wgold, zh-writing-style}

For each Dataset–ZH_METRIC pair, the following correlation scores are computed:

1. Dataset-ZH_METRIC-kendalltau
2. Dataset-ZH_METRIC-pearson
3. Dataset-ZH_METRIC-spearman
4. Dataset-ZH_METRIC-mean

where `mean` denotes the average of the three correlation coefficients. We additionally report an overall aggregated score (ALL-zh-ALL-mean), obtained by averaging correlations across both Chinese datasets and evaluation dimensions.

4. Baseline Systems

4.1. Overview

We evaluate candidate medical responses using a Large Language Model (LLM)-as-a-Judge framework. The judge receives as input: (1) the original patient query, (2) all associated medical images, (3) a set of expert-written reference responses, and (4) a candidate response to evaluate. The model outputs metric-specific scores aligned with the human evaluation protocol.

We investigate both a single-judge configuration and multi-judge ensemble variants.

4.2. MEDIQA LLM-as-a-Judge

4.2.1. Prompt Design

Full prompts are provided in Appendix.

For English samples, the judge outputs:

- `disagree_flag` $\in \{0, 1\}$
- `completeness` $\in [0.00, 1.00]$
- `factual-accuracy` $\in [0.00, 1.00]$
- `relevance` $\in [0.00, 1.00]$
- `writing-style` $\in [0.00, 1.00]$
- `overall` $\in [0.00, 1.00]$

For Chinese samples, the judge outputs:

- `factual-consistency-wgold` $\in [0.00, 1.00]$
- `writing-style` $\in [0.00, 1.00]$

The prompt includes:

- Explicit metric definitions.
- Calibration anchors (e.g., ≥ 0.85 for strong agreement, ≤ 0.30 for contradiction).
- Constraints enforcing metric independence (e.g., `overall` cannot exceed both `completeness` and `factual-accuracy`).
- Instructions to use reference responses and images as primary evidence.

4.2.2. Reference Handling

All expert reference responses are provided to the model. References are sorted by (1) frequency indicator and (2) completeness score, so that more reliable responses appear earlier. This ordering helps the judge prioritize consensus medical conclusions.

4.2.3. Continuous Scoring and Snapping

The model outputs continuous scores in $[0.00, 1.00]$. To stabilize ranking behavior and reduce numeric noise, we apply post-processing snapping to 0.05 increments:

$$\hat{s} = \text{round}\left(\frac{s}{0.05}\right) \times 0.05$$

4.2.4. Image Integration

All available images are included in the prompt as visual inputs before textual content. This ensures the judge conditions its reasoning on visual evidence, which is particularly important for woundcare cases.

4.3. Multi-Judge Ensembles

4.3.1. Temperature Ensemble

We run the LLM three times using different temperature settings: $T \in \{0.0, 0.3, 0.7\}$.

For each metric:

- Continuous scores are aggregated using the median.
- `disagree_flag` is aggregated via majority vote.

4.3.2. Persona Ensemble

We design three specialized judge prompts:

1. Diagnostic Consistency Judge
2. Clinical Safety Judge
3. Professional Communication Judge

Persona diversity encourages complementary reasoning patterns. Each judge independently produces metric scores using the same output schema. Aggregation follows the same median (continuous) and majority vote (binary) strategy.

4.4. Implementation Details

We use GPT-4o as the base model. Decoding is performed deterministically for the single-judge setting ($T = 0$). Outputs are constrained using strict JSON schema validation with automatic retry upon failure. No fine-tuning or in-context learning is applied.

5. Participating Teams & Results

The MEDIQA-EVAL 2026 shared task attracted 48 registered teams from academy and industry. Among them, five finalist teams submitted their codes and runs following the challenge rules[5]. Table 1 presents the teams that participated in the two subtasks (English and Chinese). We limited the number of submitted runs to 15 runs per team.

5.1. Code Verification

For additional validation, we required the submission of the code in addition to the models' outputs/runs. The participants shared their private codes with the organizers on GitHub following provided guidelines.

5.2. Official Results

The main results are summarized[6] in Tables 2 and 3. Our baseline systems were not included in the official ranking. On the English dataset, the SUAT-BMI team achieved the highest ALL-en-ALL-mean score (0.4816), followed by our single-LLM baseline (0.4524). On the Chinese dataset, the MedAware team obtained the highest official ALL-zh-ALL-mean score (0.2940), while our single-LLM baseline, although not part of the official ranking, achieved a higher score of 0.3271.

Our MEDIQA baseline (single-LLM) achieves strong performance on both datasets. While ensemble variants introduce additional diversity, they do not outperform the single-LLM baseline in terms of ALL-en-ALL-mean or ALL-zh-ALL-mean, suggesting that increasing output diversity alone is insufficient to improve ranking performance. Specifically, the temperature ensemble achieves ALL-en-ALL-mean and ALL-zh-ALL-mean scores of 0.4507 and 0.2767, respectively, while the persona ensemble achieves 0.4519 and 0.3032. These findings highlight the robustness of the single-LLM approach and suggest that model consistency plays a more important role than diversity in this task.

Table 4 and Table 5 present detailed results for all English (EN) metrics and correlation scores. Table 6 provides detailed results for all Chinese (ZH) metrics. Beyond overall performance, several trends emerge from the detailed results. First, performance varies across evaluation dimensions, with higher correlations observed for factual consistency and relevance, and lower scores for disagreement detection and writing quality, indicating that stylistic and contradiction-related aspects remain challenging. Second, differences are observed across correlation metrics, with Kendall's τ often yielding lower values than Pearson and Spearman correlations, reflecting its greater sensitivity to pairwise ranking differences. Despite these differences, the correlation metrics exhibit consistent trends across systems, indicating strong agreement in system rankings and reinforcing the reliability of the evaluation. Third, performance on the Chinese dataset is generally lower than on the English dataset, highlighting challenges in multilingual generalization.

Table 7 presents the correlation between common NLG metrics (ROUGE-1, BERTScore, and BLEURT, computed as both the maximum and mean across multiple references) and human judgments, alongside the MEDIQA LLM-as-a-judge baseline. On the English dataset, MEDIQA LLM-as-a-judge achieves the highest agreement with human overall ratings across all correlation measures, outperforming both traditional NLG metrics and participating systems (cf. Table 5). On the Chinese dataset, BLEURT (mean) performs best overall, with BERTScore (mean) remaining competitive, particularly for Spearman correlation. These results highlight the advantage of learned evaluators over surface-level metrics, especially in multilingual settings.

[5] Leaderboard: https://www.codabench.org/competitions/12115

[6] Full results are available at https://tinyurl.com/mediqaevalres

	Team	Affiliation	Subtasks	Paper	Code
1	SUAT-BMI	Shenzhen University of Advanced Technology, China	EN	(Peng et al., 2026)	[1]
2	SloCal-Net	University of Maribor, Slovenia & Magna Graecia University, Italy	EN/ZH	(Kocbek et al., 2026)	[2]
3	MedAware	McGill University, Canada	EN/ZH	(Hao and Liu, 2026)	[3]
4	BDI	University of Oxford, UK & University of California, USA	EN/ZH	(Xu et al., 2026)	[4]
5	hgkai26	CGI Inc., USA	EN	(Gangavarapu, 2026)	[5]

[1] https://github.com/newchat111/MEDIQA-MEDGEMMA
[2] https://github.com/pkocbek/MEDIQA-EVAL26
[3] https://github.com/Ziqi-Hao/MedAware-for-mediqa-2026
[4] https://github.com/justin13601/mediqa-eval-bdi
[5] https://github.com/harithagmu/MediQAEval2026

Table 1: MEDIQA-EVAL 2026: Participating teams, subtasks, papers, and codes.

Team	Run ID	ALL-en-ALL-mean	ALL-zh-ALL-mean
SUAT-BMI	526661	0.48156203	-
SloCal-Net	521002	0.41575853	0.25546085
MedAware	526509	0.39093232	0.25053383
BDI	526670	0.36918917	0.25244488
hgkai26	527947	0.21286092	-
MEDIQA Organizers	*526144*	*0.45239557*	*0.26694508*

Table 2: Official Results: Best Runs on the English Dataset (and their results on the Chinese Dataset)

Team	Run ID	ALL-zh-ALL-mean	ALL-en-ALL-mean
MedAware	524265	0.29400372	0.38780292
BDI	523374	0.2686269	0.33431094
SloCal-Net	524450	0.25956562	0.39715838
MEDIQA Organizers	*526255*	*0.32714034*	*0.44388561*

Table 3: Official Results: Best Runs on the Chinese Dataset (and their results on the English Dataset)

Team	Run ID	Disagree				Completeness				Factual			
		K	P	S	M	K	P	S	M	K	P	S	M
SUAT-BMI	526661	0.349	0.372	0.372	0.364	**0.416**	0.486	0.459	**0.454**	0.485	**0.553**	**0.574**	**0.537**
SloCal-Net	521002	0.250	0.261	0.266	0.259	0.392	0.423	0.409	0.408	0.475	0.523	0.518	0.505
MedAware	526509	0.117	0.118	0.118	0.118	0.316	0.397	0.336	0.350	**0.511**	0.528	0.551	0.530
BDI	526670	0.384	0.385	0.384	0.385	0.311	0.351	0.324	0.329	0.403	0.423	0.431	0.419
hgkai26	527947	0.203	0.203	0.203	0.203	0.149	0.162	0.155	0.155	0.178	0.208	0.189	0.191
MEDIQA	526144	**0.393**	**0.392**	**0.394**	**0.393**	0.398	**0.493**	**0.460**	0.450	0.414	0.504	0.483	0.467

Table 4: Official Results: Best Runs on the English Dataset (EN metrics - Part 1). K = Kendall's τ, P = Pearson correlation, S = Spearman correlation, and M = mean correlation.

Team	Run ID	Relevance				Writing				Overall				ALL
		K	P	S	M	K	P	S	M	K	P	S	M	Mean
SUAT-BMI	526661	**0.587**	**0.571**	**0.627**	**0.595**	**0.380**	0.403	**0.399**	**0.394**	0.482	0.579	0.574	0.545	**0.482**
SloCal-Net	521002	0.481	0.478	0.504	0.488	0.335	0.351	0.348	0.345	0.462	0.510	0.497	0.490	0.416
MedAware	526509	0.463	0.459	0.495	0.472	0.322	0.382	0.345	0.350	**0.492**	0.546	0.540	0.526	0.391
BDI	526670	0.346	0.317	0.357	0.340	0.293	0.316	0.301	0.304	0.418	0.456	0.444	0.439	0.369
hgkai26	527947	0.261	0.293	0.271	0.275	0.199	0.220	0.205	0.208	0.231	0.258	0.244	0.244	0.213
MEDIQA	526144	0.415	0.492	0.488	0.465	0.339	**0.446**	0.380	0.388	0.476	**0.592**	**0.583**	**0.551**	0.452

Table 5: Official Results: Best Runs on the English Dataset (EN metrics - Part 2). K = Kendall's τ, P = Pearson correlation, S = Spearman correlation, and M = mean correlation.

Team	Run ID	Factual Consistency				Writing Style				Overall
		Kendall	Pearson	Spearman	Mean	Kendall	Pearson	Spearman	Mean	
MedAware	524265	**0.471**	**0.511**	**0.511**	**0.498**	0.105	0.062	0.105	0.090	0.294
BDI	523374	0.330	0.365	0.366	0.354	0.183	0.183	0.183	0.183	0.269
SloCal-Net	524450	0.412	0.437	0.438	0.429	0.091	0.088	0.092	0.090	0.260
MEDIQA	526255	0.386	0.464	0.460	0.437	**0.193**	**0.244**	**0.215**	**0.217**	**0.327**

Table 6: Official Results: Best Runs on the Chinese Dataset (all ZH metrics)

NLG Metric	English Dataset - Overall Metric				Chinese Dataset - Factual Consistency			
	Kendall	Pearson	Spearman	Mean	Kendall	Pearson	Spearman	Mean
ROUGE-1 (max)	0.351	0.439	0.463	0.418	0.002	-0.022	0.002	-0.006
ROUGE-1 (mean)	0.372	0.501	0.498	0.457	0.002	-0.016	0.002	-0.004
BERTScore (max)	0.296	0.400	0.395	0.364	0.325	0.409	0.413	0.383
BERTScore (mean)	0.356	0.490	0.476	0.441	0.376	0.473	**0.477**	0.442
BLEURT (max)	0.333	0.439	0.440	0.404	0.336	0.425	0.427	0.396
BLEURT (mean)	0.378	0.523	0.497	0.466	0.378	**0.480**	0.476	**0.445**
MEDIQA LLM-as-a-judge	**0.476**	**0.592**	**0.583**	**0.551**	**0.386**	0.464	0.460	0.437

Table 7: Comparison of NLG metrics and LLM-as-a-judge against human evaluations on English (overall quality) and Chinese (factual consistency) datasets. MEDIQA LLM-as-a-judge achieves the strongest alignment with human overall ratings in English, while BLEURT (mean) performs best in Chinese.

6. System Descriptions

6.1. SUAT-BMI Team

The SUAT-BMI team used an ensemble evaluation framework that combines (i) few-shot LLM judging, (ii) metric-specific discriminative classifiers, and (iii) retrieval-augmented prompting. First, Qwen 30B-A3B was employed as a judge LLM in a few-shot setting, where the prompt instructs the model to compare an LLM-generated diagnosis/treatment response against multiple physician-written gold responses treated as ground truth. The judge was calibrated with labeled exemplars and then applied to rate a target sample consisting of the patient query, the candidate response, and the gold responses. Second, PubMedBERT was fine-tuned as a lightweight automatic rater. Prior to fine-tuning, MedGemma was used for zero-shot medical image captioning; the resulting caption was concatenated with the patient query and the candidate response to form an information-rich sequence. Separate PubMedBERT classifiers were trained for each evaluation metric, framing scoring as a three-way classification over 0, 0.5, 1.0, resulting in six distinct metric-specific models. Third, to mitigate hallucinations and improve evidence grounding, retrieval-augmented generation (RAG) was incorporated into the few-shot judging process. A two-stage retrieval pipeline based on LlamaIndex was used to retrieve clinical guideline passages relevant to each gold response, which were appended to the judge prompt under a dedicated "Relevant Clinical Guidelines" section.

In the final ensemble, the fine-tuned PubMedBERT was used for writing-style, RAG-enhanced few-shot judging for relevance, and standard few-shot judging for disagree-flag, factual accuracy, overall, and completeness. The system achieved an official leaderboard score of 0.481.

6.2. SloCal-Net Team

The SloCal-Net team adopted a unified pipeline consisting of evidence retrieval, context grounding, reasoning, and evaluation, with GPT-5-mini serving as the primary reasoning component. In the retrieval stage, ChatGPT Deep Research was used to automatically generate and rank 25 documents tailored to the MEDIQA-EVAL task. These documents served as an external evidence base for grounding subsequent model predictions.

Both proprietary and open-source multimodal systems were evaluated under identical experimental conditions, including GPT-5-mini, MedGemma-27B, MedGemma-1.5-4B, Qwen3-VL-32B, and Qwen3-VL-30B-Thinking. Each model was tested with and without retrieval-augmented generation (RAG) to quantify the impact of evidence grounding on evaluation performance.

In addition, the team implemented a local multimodal alternative that integrates DermLIP visual encoders with SBERT and KeyBERT for semantic alignment, followed by Qwen-based reasoning. This architecture enables clinically grounded scoring through joint visual–textual processing in a fully local pipeline.

6.3. MedAware Team

The MedAware team formulated medical response evaluation as a structured JSON prediction task

using Qwen3-VL vision–language models. All evaluation metrics were jointly predicted in a single decoding pass, conditioned on the clinical case, candidate response, gold references, and associated images. Training Procedure:
(i) *Supervised Fine-Tuning (SFT).* Qwen3-VL (8B and 32B) was fine-tuned using LoRA (rank 16, $\alpha = 32$), optimizing cross-entropy over output tokens. Model selection used 5-fold cross-validation stratified by encounter ID, evaluated via Kendall, Pearson, and Spearman correlations. The 32B LoRA model consistently outperformed smaller and fully fine-tuned variants on English metrics.
(ii) *Group Relative Policy Optimization (GRPO).* Following DeepSeek-R1, GRPO was applied on top of the SFT model. For each instance, eight outputs were sampled. The reward combined (i) exact-match accuracy with human ratings and (ii) a small formatting bonus for structured reasoning using `<think>...</think>`. GRPO optimized group-relative advantages without a separate value function.

This approach enables joint multi-metric prediction with structured outputs, combining parameter-efficient fine-tuning and reinforcement learning to better align predictions with human judgments.

6.4. BDI Team

The BDI team proposed a system-level approach that departed from the dominant fine-tuning paradigm commonly adopted in free-text medical evaluation. Instead of training or adapting task-specific language models, the team implemented an agent-based evaluation framework that orchestrated multiple existing components via API access to an off-the-shelf language model.

The system was built around a ReAct-style agent that performed structured reasoning guided by task-specific prompts. The agent integrated multimodal retrieval of similar clinical encounters, auxiliary machine learning–based scoring signals, AI-generated VQA references, and optional image augmentation tools, without any additional pre-training or fine-tuning. This design investigated the feasibility of modular, compositional agents as a lightweight and computationally efficient alternative for free-text medical QA evaluation, reflecting broader trends toward agentic systems in applied NLP and multimodal reasoning.

6.5. hgkai26 Team

The hgkai26 team implemented a multimodal LLM-as-a-Judge evaluation framework built on GPT-4o within the OpenAI Evals infrastructure to systematically assess system-generated responses for dermatology and wound care queries. Rather than relying on manual annotation or task-specific classifier training, the framework leveraged a rubric-guided large language model to perform structured, criteria-based grading of candidate answers.

Each evaluation instance consisted of a user query, an associated clinical image, and a candidate response. For every query, three candidate responses were independently evaluated using standardized prompts that encoded predefined scoring rubrics. The judge model assigned dimension-level scores across key quality attributes—including completeness, factual accuracy, relevance, and writing quality—using a discrete scoring scale of 0, 0.5, 1. In addition to these granular assessments, the system produced an aggregate overall score and a binary disagreement indicator to capture potential inconsistencies in evaluation outcomes. The evaluation pipeline was fully automated, generating structured quantitative outputs that enabled consistent, reproducible comparison of candidate responses. This design facilitated scalable benchmarking of multimodal medical question-answering systems by operationalizing helpfulness and response quality through rubric-based, model-driven assessment.

7. Conclusion

In this paper, we introduced the MEDIQA-EVAL 2026 shared task on evaluation metrics for medical multimodal question answering, motivated by the growing gap between rapid advances in multimodal clinical QA systems and the limited reliability of existing evaluation methods. Prior work has shown that existing automatic metrics, whether reference-based or LLM-based, do not consistently align with clinician judgments, particularly in specialized medical settings. This gap is especially evident in open-ended MMQA, where quality depends on factual accuracy, visual relevance, completeness, and overall coherence rather than surface similarity alone. By shifting the focus to evaluating the evaluators, our shared task establishes a clinician-grounded benchmark for automatic metric assessment in medical multimodal QA.

The participating systems span diverse methodological directions, including fine-tuned vision-language models, reinforcement learning, retrieval-augmented judging, metric-specific classifiers, agent-based pipelines, and prompting-based LLM-as-a-Judge frameworks. Together, the results provide a comparative overview of current approaches to multilingual and multimodal clinical evaluation. Through clinician-rated data and multidimensional quality criteria, this shared task offers a foundation for developing more reliable and clinically meaningful evaluation methods. Future work will focus on expanding the datasets by increasing the number of responses and ratings for English and Chinese.

Appendix: Prompt Templates

MEDIQA LLM-as-a-Judge: English Prompt

```
You are given a patient query, relevant medical images, a list of reference responses, and a candidate response. The reference responses were written by practicing medical doctors.
Please evaluate the candidate response according to the following metrics, using the same criteria as a practicing medical doctor.
Metrics: - disagree_flag (0 or 1): 1 if the candidate response clearly disagrees with the expert reference responses on key medical facts; 0 otherwise. - completeness (0.00 to 1.00): How completely the response answers the patient's question. - factual-accuracy (0.00 to 1.00): Whether the medical information is factually accurate. - relevance (0.00 to 1.00): Whether the response is relevant to the patient's question and images. - writing-style (0.00 to 1.00): Clarity, professionalism, and appropriateness of tone.
- overall (0.00 to 1.00): Overall quality considering completeness and factual accuracy primarily.
Calibration guidance: - Responses that closely match most reference responses in diagnosis and management should receive scores ≥ 0.85 for completeness, factual-accuracy, and overall. - Responses that partially match references should receive scores around 0.40-0.70.
- Responses that contradict references or omit key information should receive scores ≤ 0.30.
Constraints: - Score each metric independently. - A high writing-style score must not increase factual-accuracy or completeness.
- overall must not be higher than both completeness and factual-accuracy. - If either completeness or factual-accuracy ≤ 0.50, overall should also be ≤ 0.50.
Use the reference responses and the medical images as the primary evidence for your evaluation.
Output requirements: - For disagree_flag: output 0 or 1 only. - For all other metrics: output a continuous score in [0.00, 1.00] with two decimals.
Return ONLY a valid JSON object with exactly the following keys: disagree_flag, completeness, factual-accuracy, relevance, writing-style, overall All values must be numeric.
```

MEDIQA LLM-as-a-Judge: Chinese Prompt

```
你将看到一个患者问题、相关医学图像、若干参考回答，以及一个候选回答。参考回答由医学专家提供。
请按照中国医学院受训的医学专家的评估标准，对候选回答进行评分。
评分指标：- factual-consistency-wgold (0.00 到 1.00)：表示候选回答在医学事实上与金标准参考回答的一致程度。1.00 表示在诊断、判断或处理建议上与参考回答完全一致；0.00 表示与参考回答存在明显矛盾或医学错误；中间值表示部分一致或遗漏部分要点。
- writing-style (0.00 到 1.00)：表示回答在中文语言表达上的专业性、清晰度和得体程度。1.00 表示语言自然、专业，符合中文医学表达习惯；0.00 表示表达不清晰、不专业或明显不合适；中间值表示基本通顺但略显生硬或不够规范。
重要说明：- factual-consistency-wgold 只评估医学事实是否与参考回答一致，不要考虑语言是否流畅。- writing-style 只评估语言表达质量，不要考虑医学内容是否正确。- 两个评分必须相互独立，不要相互影响。
评分参考：- 若候选回答在关键医学结论上与参考回答高度一致，factual-consistency-wgold 通常应 ≥ 0.85。- 若仅部分一致或遗漏部分要点，通常在 0.40-0.70 之间。- 若存在关键矛盾或明显错误，通常 ≤ 0.30。
医学图像应作为判断医学事实一致性的重要依据。
输出要求：- 所有评分必须是 [0.00, 1.00] 范围内的连续数值（保留两位小数）。
请仅返回一个 JSON 对象，且必须包含以下键：factual-consistency-wgold, writing-style 所有值必须是数值。
```


Acknowledgements

We would like to thank Thomas Lin from Microsoft Health AI and the ClinicalNLP organizers for their support for the shared task. We also thank the ClinicalNLP reviewers, our annotation team, and the participating teams who contributed to the success of the shared task through their interesting approaches and strong engagement.

Limitations

This shared task does not cover all possible methods for multimodal question answering evaluation. In addition, the test set is limited in both size and domain coverage, focusing primarily on dermatology. As a result, the findings may not generalize to other medical specialties or broader clinical settings. Further validation on larger and more diverse datasets is necessary to assess the robustness and generalizability of the best-performing approaches.

8. Bibliographical References

Asma Ben Abacha, Vivek V. Datla, Sadid A. Hasan, Dina Demner-Fushman, and Henning Müller. 2020. Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 1081–1088. CEUR-WS.org.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An investigation of evaluation methods in automatic medical note generation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pages 2575–2588. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6556–6576. Association for Computational Linguistics.

Haritha Gangavarapu. 2026. hgkai26 at mediqa-eval 2026: Automatically evaluating the accuracy of ai-generated responses in visually driven specialties. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026*, Palma, Mallorca, Spain. European Language Resources Association.

Ziqi Hao and Pengbo Liu. 2026. Medaware at mediqa-eval 2026: Vision-language model fine-tuning with logprob-based score calibration for medical response evaluation. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026*, Palma, Mallorca, Spain. European Language Resources Association.

Xuehai He, Yichen Zhang, Luntian Mou, Eric P. Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *CoRR*, abs/2003.10286.

Primoz Kocbek, Valentina Carbonari, Pierangelo Veltri, Pietro Hiram Guzzi, and Gregor Stiglic. 2026. Slocal-net at mediqa-eval 2026: Investigating the impact of reasoning and external context on medical answer grading. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026*, Palma, Mallorca, Spain. European Language Resources Association.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

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.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Xinzhe Peng, Liyuan E, Kun Feng, Jielin Li, Yuxuan Tang, and Zhao Li. 2026. Suat-bmi at mediqa-eval 2026: An ensemble approach to language models as judges for automatic rating of medical responses. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026*, Palma, Mallorca, Spain. European Language Resources Association.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Annual Meeting of the Association for Computational Linguistics*.

Justin Xu, Zizheng Zhang, Augustine Luk, Benjamin Khong, Haochen Cui, Samuel Hwang, Alyssa Pradhan, Kevin Yuan, and David W. Eyre. 2026. Bdi at mediqa-eval 2026: A react-style multimodal agent for fine-grained medical response

assessment. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026*, Palma, Mallorca, Spain. European Language Resources Association.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2026. MORQA: benchmarking evaluation metrics for medical open-ended question answering. In *LREC, Mallorca, Spain, 13-15 May 2026*.

Wen-wai Yim, Asma Ben Abacha, Noel Codella, Roberto A. Novoa, and Josep Malvehy. 2025a. Overview of the MEDIQA-MAGIC task at imageclef 2025: Multimodal and generative telemedicine in dermatology. In *Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025*, volume 4038 of *CEUR Workshop Proceedings*, pages 2233–2240. CEUR-WS.org.

Wen-wai Yim, Asma Ben Abacha, Robert Doerning, Chia-Yu Chen, Jiaying Xu, Anita Subbarao, Zixuan Yu, Fei Xia, M. Kennedy Hall, and Meliha Yetisgen. 2025b. Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care. *J. Biomed. Informatics*, 170:104888.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen, and Martin Krallinger. 2024a. Overview of the MEDIQA-M3G 2024 shared task on multilingual multimodal medical answer generation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop, ClinicalNLP@NAACL 2024, Mexico City, Mexico, June 21, 2024*, pages 581–589. Association for Computational Linguistics.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, and Fei Xia. 2024b. Overview of the MEDIQA-MAGIC task at imageclef 2024: Multimodal and generative telemedicine in dermatology. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 1456–1462. CEUR-WS.org.

Wen-wai Yim, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2025c. Overview of the MEDIQA-WV 2025 shared task on woundcare visual question answering. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 17–21, Virtual. Association for Computational Linguistics.

Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2024c. Dermavqa: A multilingual visual question answering dataset for dermatology. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part V*, volume 15005 of *Lecture Notes in Computer Science*, pages 209–219. Springer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

SUAT-BMI at MEDIQA-EVAL 2026: An Ensemble Approach to Language Models as Judges for Automatic Rating of Medical Responses

Xinzhe Peng*, Liyuan E*, Kun Feng, Jielin Li, Yuxuan Tang, Zhao Li
Shenzhen University of Advanced Technology
lizhao@suat-sz.edu.cn

Abstract

The MEDIQA-EVAL 2026 shared task focuses on developing automatic evaluation metrics for LLM-generated responses in dermatology and wound care. While LLMs have shown promise as judge models, the reliability of these metrics remains underexplored. In this work, we study how well judge models can approximate human expert ratings across clinical evaluation criteria. We evaluate multiple approaches, including few-shot prompting, BERT fine-tuning, and retrieval-augmented generation (RAG), and combine them in an ensemble framework. Our method achieves a correlation score of 0.481, ranking first among 41 participating teams. Our results provide insight into the reliability of LLM-based evaluation metrics and highlight their potential for scalable clinical assessment.

Keywords: Medical Question Answering, LLM, BERT, Retrieval Augmented Generation

1. Introduction

Evaluating LLM-generated responses to clinical questions remains a significant challenge due to the complexity of medical contexts and the need for nuanced semantic judgment. Such evaluations are inherently multidimensional, requiring assessment of factual accuracy, completeness, relevance, and writing quality. While expert review by trained professionals provides high-quality assessments, it is costly and difficult to scale, motivating the development of automated evaluation methods.

Recent work has increasingly explored the use of large language models (LLMs) as judge models to automatically evaluate generated outputs. These approaches rely on predefined evaluation metrics, yet it remains unclear whether such metrics, when used as criteria by LLMs, produce ratings that are consistent with human expert judgment. Understanding the reliability of these metrics is therefore a critical question.

In this work, we investigate the extent to which LLM-based judge models can approximate expert evaluations under established clinical assessment criteria. Using the MEDIQA-EVAL 2026 shared task dataset, which includes patient queries, LLM-generated responses, expert reference answers, and fine-grained ratings provided by dermatology and wound care specialists (Yim et al., 2024, 2025a,b), we frame the problem as predicting expert-assigned scores across multiple evaluation dimensions.

To study this, we propose an ensemble-based judge model that integrates multiple approaches, including supervised fine-tuning and few-shot prompting with models such as Qwen3-30B-A3B and PubMedBERT, along with retrieval-augmented generation (RAG). By selecting model configurations based on development set performance, our system aims to maximize alignment with expert ratings, achieving a final official score of 0.481.

Our results show that a carefully constructed ensemble of judge models can approximate human evaluation to a meaningful degree, while also revealing variability in how well different metrics are captured. These findings suggest that while LLM-based evaluators hold promise for scalable clinical assessment, their reliability remains dependent on both model design and the evaluation criteria being applied. This work provides insight into the strengths and limitations of judge models and offers a foundation for developing more robust automated evaluation methods in specialized domains.

2. Related Work

Prior research has shown that LLMs can generate unreliable responses in medical question-answering settings, where hallucinations and factual inaccuracies remain significant challenges (Reichenpfader et al., 2024; Yu et al., 2025). These limitations raise concerns about the safe deployment of LLMs in healthcare applications and highlight the need for robust evaluation methodologies.

To address this issue, a growing body of work has focused on systematically evaluating LLM-generated responses in medical domains. One approach relies on assessment by qualified biomedical experts. For instance, the CLEVER framework provides a structured methodology for trained medical professionals to evaluate LLM outputs (Kocaman et al., 2025). While expert evaluation is considered the gold standard, it is inherently time-

*These authors contributed equally.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 12–18
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

consuming and costly, limiting its scalability.

Alternatively, automatic evaluation metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) have been widely adopted to quantify the quality of generated text. However, these metrics primarily measure surface-level similarity and often fail to capture the nuanced clinical reasoning and domain-specific knowledge required in medical question answering. As a result, their effectiveness in evaluating medical responses remains limited.

Motivated by these shortcomings, recent work has explored the use of LLM-based evaluators under the "LLM-as-a-judge" paradigm. For example, Medical Eval Sphere introduces an open-source benchmark for long-form medical question answering based on real-world patient queries and expert-annotated reference responses (Hosseini et al., 2024). Similarly, MedThink-Bench presents a dataset of complex medical questions and incorporates an LLM-based judge to evaluate reasoning quality (Zhou et al., 2025). These efforts demonstrate the potential of automated, model-based evaluation frameworks in capturing more nuanced aspects of response quality.

Despite these advances, limited work has systematically examined how to train judge models that can reliably approximate expert evaluations, particularly with respect to the consistency of the underlying evaluation metrics. The impact of different training strategies and model combinations on metric reliability remains underexplored. In this work, we address this gap by leveraging the MEDIQA-EVAL dataset to study the performance of judge models across diverse learning paradigms. We further propose an ensemble-based approach and show that combining multiple models leads to more robust alignment with expert ratings than any single method.

3. Dataset and Task Description

This section outlines the dataset and the evaluation objectives of the MEDIQA-EVAL 2026 shared task. It details the composition and annotation schema of the dataset, with a formal definition of the evaluation metrics used to measure the effect between our system's predictions and expert judgments.

3.1. Dataset

The MEDIQA-EVAL 2026 shared task dataset is based on two multimodal medical visual question answering (VQA) collections: **DermaVQA** (dermatology cases from the MEDIQA-M3G challenge) and **WoundcareVQA** (wound care cases from the MEDIQA-WV 2025 challenge). These datasets consist of online medical posts, which include both textual descriptions and clinical images, paired with responses provided by medical professionals. Each clinical case is provided with gold-standard expert responses along with three distinct system-generated candidate responses.

Annotation Schema and Metrics To construct the ground-truth ratings, the system-generated responses were evaluated by human experts. There are 6 metric to consider.

- **Disagree_Flag:** if human experts disagree with an answer.
- **Factual-Accuracy:** if the answer contains all necessary information required to give to the patient based on the question present, rate from 0 and 1.
- **Relevance:** if everything in the response is medically accurate, rate from 0, 0.5, 1.0.
- **Completeness:** if the response covers all aspects of the user's query, rate from 0, 0.5, 1.0.
- **Writing-Style:** evaluates the linguistic quality, rate from 0, 0.5, 1.0.
- **Overall:** the overall score of the response, rate from 0, 0.5, 1.0.

Our work exclusively focuses on the **English (EN) sub-task**, utilizing the 2,898 data in the development set and 3,474 in the test set. The development set is the set we utilize to train our model, and the test set is the final evaluation set that computes the score on the leaderboard. The participants could not access the test set during the competition, though it was released after the competition for verification and reference.

3.2. Task

The MEDIQA-EVAL 2026 shared task evaluates the performance of the rating model using correlation with expert ratings. The objective is to produce predicted ratings that are the most correlated with expert-provided ratings. The final score is computed as the mean of Kendall, Spearman, and Pearson correlations according to the following formulas:

Define n to be the number of samples. $\hat{Y}_{1:n} \in R$ as the predicted scores and $Y_{1:n} \in R$ as the labeled scores for each metric.

Let

$$y_i, y_j \in Y, \quad \hat{y}_i, \hat{y}_j \in \hat{Y}, \quad i, j \in \{1, 2, \ldots, n\},\ i < j$$

Then, estimations of the correlations are:

$$\hat{\rho}_{pearson} = \frac{Cov(Y,\hat{Y})}{\hat{\sigma}_Y\, \hat{\sigma}_{\hat{Y}}}$$

$$\hat{\rho}_{spearman} = \frac{Cov(rank(Y), rank(\hat{Y}))}{\hat{\sigma}_{rank(Y)} \hat{\sigma}_{rank(\hat{Y})}}$$

$$\hat{\rho}_{kendall} = \frac{2}{n(n-1)} \sum_{i<j} sgn(y_i - y_j)\, sgn(\hat{y}_i - \hat{y}_j)$$

The final correlation is calculated as the mean of the three correlations:
$\hat{\rho} = \frac{\hat{\rho}_{pearson}+\hat{\rho}_{spearman}+\hat{\rho}_{kendall}}{3}$

4. Methodology

4.1. Few-Shot Learning

Our Few-Shot Learning method aims to help the judge LLM to understand the underlying ratings rules of the six metrics. We select Qwen3-30B-A3B as our base language model for its inference speed and text-only task performance(QwenTeam, 2025). There are 483 English samples in the development set of the task. To implement few-shot learning strategy, we perform 7 separate runs. In each run, we bootstrap 20 samples and feed them to the model. For each sample, the final output is computed by averaging the predictions from all 7 runs. The approach is shown in Fig 1.

4.1.1. Prompt Design

Prompt

You are an expert in dermatology and wound care. Think silently if needed .Your job is to give accurate ratings of the response generated by an LLM, which includes a diagnosis and a treatment. The response is an answer to a question from a patient asking about a disease that he has. You will be supplied with gold responses given by professional doctors. You should rate the LLM's response by comparing it very carefully with the gold responses and treat them as ground truth.

...<rating rules>...

I will give you a few examples with ratings. After that, you should provide your ratings to a sample, which contains only a query, a response and gold responses.

...<shots>...

The sample you need to rate:
<patient's query>
<LLM's response to the patient's query>
<gold responses to the patient's query>

For each sample, we construct our prompt with information containing the user's query, the LLM's response to the query, the gold responses, official rating rules presented by the MEDIQA-EVAL 2026 organizers, and the shots. We specifically instruct the rating LLM to compare carefully with the gold responses and deduce the underlying rating rules from the shots containing expert-labeled ratings. Using gold responses is a limitation of our approach, as it prevents direct application to datasets outside MEDIQA-EVAL. This design is intended to efficiently rate responses within this specific dataset.

4.1.2. Bootstrapped Few-Shot Learning

Although the rating LLM yields a decent result after few-shot learning with 20 shots, the rating LLM is still blind to other 483 training data . With only 20 shots, the rating LLM fails to see all possible combinations of metric scores. Therefore, we construct 7 conversations for each sample where each prompt contains bootstrapped shots from the training data, and the final output for the sample is the average of 7 LLM outputs from 7 conversations. We found that by bootstrapping shots multiple times and aggregating the final output by the mean consistently outperforms the model where the LLM is provided the shots for only once. We hypothesize that the ensemble of the 7 rating LLMs learns a better approximation of the distribution by sampling shots with replacement.

4.2. Fine-tuning BERT

4.2.1. Model Selection

PubMedBERT was selected as the base text encoder considering that it was pretrained on biomedical literature. Due to the same reason, *MedGemma27b-it* was employed for zero-shot image captioning. This effectively translates visual clinical information into textual descriptions, enabling the text-only BERT model to process multimodal information.

4.2.2. Targeted Upsampling

To decrease the severe class imbalance in the WoundcareVQA subset , a targeted upsampling mechanism was applied. Minority class instances (scores 0.0 and 0.5) were oversampled in the training folds to prevent the majority-class collapse.

4.2.3. Metric-Specific Input Construction

To accurately predict the six distinct evaluation metrics, it has naturally occurred that different metrics need different amount of information. Thus a metric-specific strategy was engineered.

For metrics like *writing-style*, *completeness* dependent primarily on linguistic quality, inputs were restricted strictly to the LLM response. The input sequence is constructed as the following: `[LLM Response] <response>`.

Conversely, for metrics like *relevance*, *factual-accuracy*, we constructed the inputs by concatenating the MedGemma27b-generated caption, query, LLM response. The input sequence is constructed as the following: `[IMG] <caption> [QUERY] <query> [LLM Response] <response>`.

4.2.4. Training and Inference Strategy

To fine-tune BERT for score generation, we adopt a tunable linear projection layer and apply a sigmoid function at the end to output a score from 0 to 1, as shown in Fig 1.

To maximize data utilization and prevent overfitting, a 5-fold cross-validation framework was employed. Models were optimized using cross-entropy loss over the three target classes. During inference, an internal probability averaging mechanism was utilized. This aggregation produces robust baseline predictions for the final system ensemble.

4.3. Retrieval Augmented Generation

Knowledge Base Construction: To ensure the factual accuracy of the evaluation, we constructed a domain-specific knowledge base utilizing 43 clinical guidelines focused on dermatology and wound care (World Health Organization, 2026; National Institute for Health and Care Excellence, 2026; NHS England, 2023; American Academy of Dermatology, 2026; Wound, Ostomy and Continence Nurses Society, 2026; Infectious Diseases Society of America, 2026; Wound Healing Society, 2026; American Physical Therapy Association, 2026). The majority of documents were systematically curated from authoritative international and professional repositories. They were retrieved from the official websites of leading medical institutions. Additionally, we incorporated evidence-based guidelines from specialized professional societies. We employed a sliding window strategy for text chunking, with a chunk size of 1024 tokens and an overlap of 100 tokens to preserve semantic continuity across boundaries.

Embedding and Indexing: We implemented a two-stage retrieval pipeline using the LlamaIndex framework (Liu, 2022) to retrieve relevant clinical evidence for a given gold response. The chunked text data was mapped into a high-dimensional vector space using the bge-m3 embedding model (Chen et al., 2024a). bge-m3 was selected for its robust performance in multilingual retrieval and support for dense retrieval tasks.

Retrieval and Reranking: For each patient query, we first performed an approximate nearest neighbor search to retrieve the top-10 candidate chunks. To further improve relevance and filter out noise, we applied a cross-encoder reranking step using bge-reranker-large (Xiao et al., 2023). This model re-scores the candidate chunks based on their fine-grained interaction with the query. The top-ranked chunks were then concatenated the top-3 chunks to form the context block denoted as <Relevant Clinical Guidelines> in the prompt.

4.4. Ensemble

As indicated by preliminary experiments on the development set, system performance varies across evaluation criteria, and no single modeling approach consistently achieves optimal performance across all metrics. The independent methods considered in this work—supervised BERT fine-tuning, Few-Shot LLM prompting, and Retrieval-Augmented Generation (RAG)—exhibit complementary strengths under different evaluation settings.

To leverage these complementary properties, we adopt an ensemble strategy for the final submission. Instead of employing a uniform or weighted aggregation scheme, we implement a dataset- and metric-specific selection mechanism. Concretely, for each combination of dataset and evaluation metric, the final prediction is derived exclusively from the model that achieves the highest correlation on the set. This design choice is motivated by empirical observations that model performance is highly dependent on both the task formulation and evaluation dimension.

- **Supervised BERT for Writing-Style:** The fine-tuned BERT model was exclusively selected for *writing-style* across both datasets. This confirms that supervised learning on domain-specific data remains the most effective method for capturing the stylistic consistency of medical responses.
- **RAG for Completeness:** The Retrieval-Augmented Generation (RAG) framework demonstrated superior performance on *relevance* for both datasets. This suggests that retrieving similar historical cases aids significantly in establishing the semantic alignment between the user query and the candidate response.
- **Few-Shot LLM for Factual Accuracy, Overall Ratings, Relevance and Disagree Flag:** The Large Language Model (LLM) proved most robust for high-level reasoning tasks, serving as the primary predictor for *factual-accuracy* and *overall* ratings. This underscores the necessity of leveraging the broad medical knowledge and zero-shot reasoning capabilities of LLMs for complex factual verification.

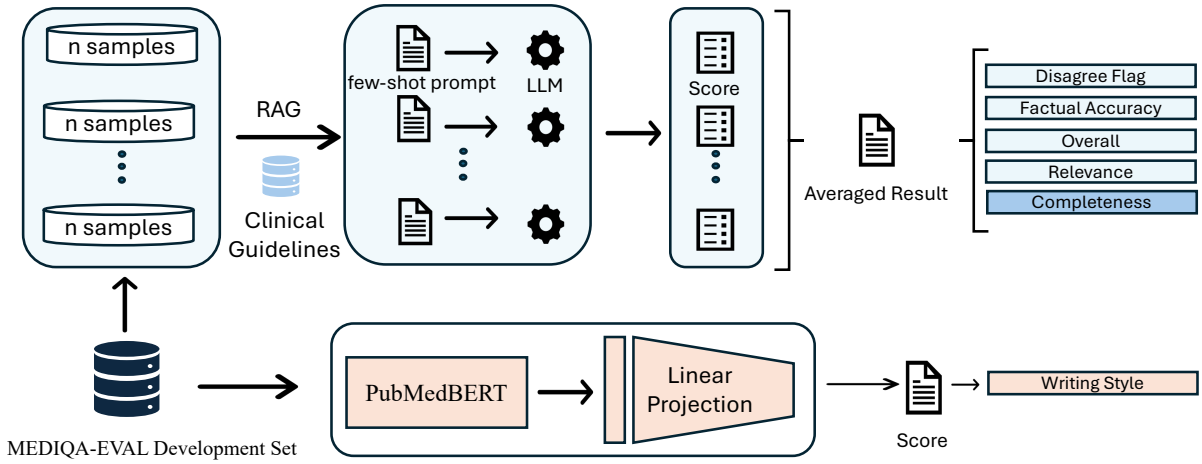


Figure 1: The proposed ensemble architecture for medical response evaluation.

Model	Disagree	Complete	Factual	Relevance	Style	Overall
Few-Shot	**0.48**	**0.39***	**0.44**	**0.57**	0.47	**0.47**
BERT	0.23	0.33	0.29	0.40	**0.52**	0.28
RAG	0.23	**0.38***	0.27	0.28	0.25	0.39
Ensemble	**0.48**	**0.39**	**0.44**	**0.57**	**0.52**	**0.47**

Table 1: Empirical result of each model on the development set.
*It can reasonably be assumed that RAG and few-shot learning perform equally well on this task. We chose RAG for metric 'completeness' on the test set.

By systematically exploiting these complementary strengths—fine-tuning for style, retrieval for relevance, and LLM reasoning for accuracy—our final system underscores the necessity of constructing an ensemble for complex medical evaluation tasks.

4.5. Results

The performance of our final ensemble system was evaluated on both the official MEDIQA-EVAL 2026 English development set and test set. The system, which integrates the strengths of supervised fine-tuning, retrieval-augmented generation, and few-shot learning, achieved a global aggregate mean correlation of **0.4816** (`ALL-en-ALL-mean`) on the test data.(Asma Ben Abacha, 2026)

Table 1 and Table 2 represent the detailed scores on development set and test set, respectively. The system demonstrated exceptional performance on the six metrics. This suggests that the ensemble strategy effectively synthesizes multimodal context and external knowledge to verify the semantic alignment and factual correctness of medical responses. It also indicates that the ensemble generalizes to data outside of the development set.

4.6. Error Analysis

To study the weakness of our model, we look at predictions that have a difference between the label larger than 0.5 on the development set. We found that poor predictions prevail at disagree flag and factual accuracy. A simple statistic showing where errors occur is shown in Table 3. These two metrics require solid knowledge of medical care and strong reasoning skills to detect the flaw in the response compared to other metrics. Further investigation is needed to understand the reason behind the large discrepancy of disagreeing a response between the judge model and human evaluators.

5. Conclusion

In this work, we show that an ensemble of few-shot learning, fine-tuned PubMedBERT, and RAG-enhanced prompting outperforms any individual model, with each component contributing distinct strengths across evaluation metrics. Specifically, fine-tuned PubMedBERT performs well on simpler classification tasks such as writing style, while LLM-based few-shot prompting is effective for metrics requiring medical reasoning, such as factual correctness. Incorporating retrieval-augmented information further improves performance on completeness. The consistent performance of our ensemble across both development and test sets suggests its robustness and generalizability.

Despite achieving strong overall results, the model shows weaker correlation on metrics such as disagree flag and writing style, and struggles with

Model	Disagree	Complete	Factual	Relevance	Style	Overall
Few-Shot	**0.36**	0.45	**0.54**	**0.60**	0.40	**0.55**
BERT	0.20	0.38	0.30	0.33	**0.40**	0.36
RAG	0.34	**0.45**	0.52	0.37	0.37	0.52
Ensemble	**0.36**	**0.45**	**0.54**	**0.60**	**0.40**	**0.55**

Table 2: Experiment on the contribution of each model on the test set, consistent with the findings on the development set

Metric	Count
Disagree_flag	70
Factual-Accuracy	12
Completeness	8
Overall	5
Relevance	2
Writing-Style	0

Table 3: Distribution of metrics for which error is higher than 0.5

aspects of factual accuracy. These limitations highlight the need for future work on improving medical reasoning capabilities and developing more reliable evaluation frameworks that better align with expert judgment.

Overall, our findings indicate that a carefully designed judge model can achieve meaningful alignment with expert ratings, though performance remains metric-dependent. Without appropriate design choices, individual methods may produce evaluations that are inconsistent with those of human experts.

6. Bibliographical References

American Academy of Dermatology. 2026. Aad guidelines of care. [EB/OL]. [2026-02-03].

American Physical Therapy Association. 2026. Apta clinical practice guidelines. [EB/OL]. [2026-02-03].

Wen-wai Yim Asma Ben Abacha. 2026. Overview of the mediqa-eval 2026 shared task on evaluation metrics in medical multimodal question answering. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

M. L. Chen et al. 2024b. Performance and risk of harm of a large language model on dermatology continuing medical education questions. *Journal of Investigative Dermatology*, 144(8):S25.

Xiaolan Chen, Jiayang Xiang, Shanfu Lu, Yexin Liu, Mingguang He, and Danli Shi. 2025. Evaluating large language models and agents in healthcare: key challenges in clinical applications. *Intelligent Medicine*, 5(2):151–163.

Luca Corradini, Giulia Marcaccini, Isha Seth, Warren M. Rozen, Chiara Biagiotti, Roberto Cuomo, and Francesco R. Giardino. 2025. Ai vs. md: Benchmarking chatgpt and gemini for complex wound management. *Journal of Clinical Medicine*, 14(24):8825.

Pedram Hosseini, Jessica M Sin, Bing Ren, Bryceton G Thomas, Elnaz Nouri, Ali Farahanchi, and Saeed Hassanpour. 2024. A benchmark for long-form medical question answering. *arXiv preprint arXiv:2411.09834*.

Infectious Diseases Society of America. 2026. Idsa practice guidelines. [EB/OL]. [2026-02-03].

Veysel Kocaman, Murat Kaya, Adrian Feier, and David Talby. 2025. Clinical large language model evaluation by expert review (clever): Framework development and validation. *JMIR AI*, 4:e72153.

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.

Jerry Liu. 2022. LlamaIndex.

National Institute for Health and Care Excellence. 2026. Nice guidelines: Evidence-based recommendations for health and care. [EB/OL]. [2026-02-03].

NHS England. 2023. Wound care information standard. [EB/OL]. [2026-02-03].

QwenTeam. 2025. Qwen3 technical report.

Daniel Reichenpfader, Peter Rösslhuemer, and Kerstin Denecke. 2024. Large language model-based evaluation of medical question answering systems: Algorithm development and case study. *Studies in Health Technology and Informatics*, 313:22–27.

World Health Organization. 2026. World health organization: Guidelines and technical documents. [EB/OL]. [2026-02-03].

Wound Healing Society. 2026. Wound healing society clinical resources and guidelines. [EB/OL]. [2026-02-03].

Wound, Ostomy and Continence Nurses Society. 2026. Wocn society: Clinical guidelines and resources. [EB/OL]. [2026-02-03].

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Wen-wai Yim, Asma Ben Abacha, Robert Doerning, Chia-Yu Chen, Jiaying Xu, Anita Subbarao, Zixuan Yu, Fei Xia, M. Kennedy Hall, and Meliha Yetisgen. 2025a. Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care. *Journal of Biomedical Informatics*, 170:104888.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2025b. MORQA: benchmarking evaluation metrics for medical open-ended question answering. volume abs/2509.12405.

Ww. Yim, Y. Fu, Z. Sun, A. B. Abacha, M. Yetisgen, and F. Xia. 2024. Dermavqa: A multilingual visual question answering dataset for dermatology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15005 of *Lecture Notes in Computer Science*, Cham. Springer.

E Yu, X Chu, W Zhang, X Meng, Y Yang, X Ji, and C Wu. 2025. Large language models in medicine: Applications, challenges, and future directions. *International Journal of Medical Sciences*, 22(11):2792–2801.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Sheng Zhou, Wei Xie, Jian Li, Zhi Zhan, Ming Song, Hui Yang, Carlos Espinoza, Laura Welton, Xiang Mai, Yan Jin, Zhen Xu, Young-Hoon Chung, Yi Xing, Ming-Hsiu Tsai, Eric Schaffer, Yi Shi, Nan Liu, Zhi Liu, and Rong Zhang. 2025. Automating expert-level medical reasoning evaluation of large language models. *NPJ Digital Medicine*, 9(1):34.

Overview of the MEDIQA-SYNUR 2026 Shared Task on Observation Extraction from Nurse Dictations

George Michalopoulos, Jean-Philippe Corbeil, Cari Bader
Nathan Bodenstab and Asma Ben Abacha
Microsoft Healthcare & Life Sciences
{georgemi, jcorbeil, caribader, nbodenstab, abenabacha}@microsoft.com

Abstract

Hospital nurses spend a significant portion of their shifts performing manual data entry tasks. An automatic solution for extracting medical information from nurse dictations into large spreadsheet ontology (flowsheet) could reduce the documentation burden of nurses and alleviate nurse burnout. We introduce the MEDIQA-SYNUR shared task, the first challenge on extracting and normalizing clinical observations from nurse dictations and mapping them to a large ontology of clinical concepts. 13 teams participated in the challenge and experimented with a broad range of approaches. In this paper, we describe the MEDIQA-SYNUR task, the datasets, and the participant's results and solutions.

Keywords: nursing shared task, observation extraction, ontology mapping

1. Introduction

Nurses currently manually enter information into the electronic health record (EHR) using flowsheets, which are similar to spreadsheets. Each flowsheet can have multiple tabs and hundreds or even thousands of rows, with each row representing a specific data point such as temperature, heart rate, etc. Working with flowsheets is a time-consuming process, since the nurse must first navigate to the appropriate tab in the flowsheet, identify the row or groups of rows corresponding to the assessment that is being documented, then manually enter the information cell by cell. Due to the complexity of the flowsheet, nurses often cannot document during rounds and instead delay this task until breaks, leading to delayed documentation and potential accuracy issues.

Surveys show growing interest in NLP for both nursing notes (Mitha et al., 2023) and wider nursing tasks (Panchal and Thakur, 2024). Yet since nurses still chart mainly in electronic health record (EHR) flowsheets, any automation must mesh with that structure.

LLMs can already translate clinical narratives into structured variables with little or no task-specific training (Ling et al.; Dagdelen et al., 2024). Yet their edge over fine-tuned encoder models remains debated (Gutiérrez et al., 2022; Ling et al., 2023). Consequently, no existing system simultaneously (i) ingests nurse dictations, (ii) conditions on the local flowsheet context, and (iii) outputs EHR-ready observations.

The MEDIQA-SYNUR shared task tackles the problem of extracting and normalizing clinical observations from nurse dictations and mapping them to a large ontology of clinical concepts (flowsheet), as described in Figure 1.

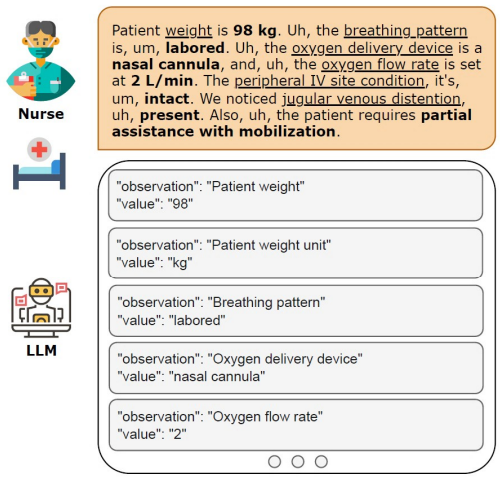


Figure 1: The observation extraction from nurse dictations task takes a nurse dictations and a local flowsheet and extracts EHR-ready observations

In this paper, we present the task and the datasets in Sections 2 and 3. In Section 4, we present the evaluation metric used in the shared task. Section 5, describes the approaches of the participating teams, and in Section 6 we discuss the final insights from the official challenge results.

2. Task Description

The MEDIQA-SYNUR 2026 shared task[1] addresses the problem of extracting clinical observations from nurse dictations.

[1] https://sites.google.com/view/mediqa2026/mediqa-synur

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 19–26
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

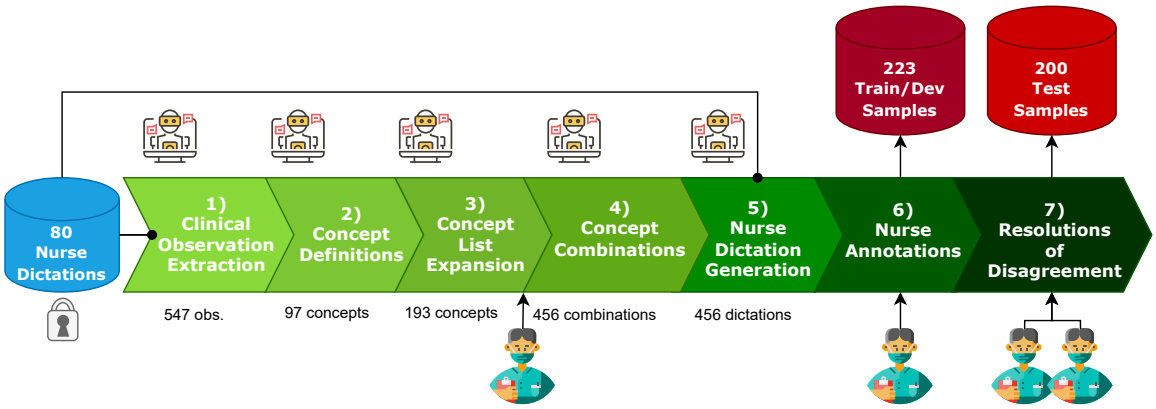


Figure 2: SYNUR Dataset Creation Pipeline for generating realistic, synthetic nurse dictations with expert nurse annotations, comprising six steps: starting from 80 human-verified, fabricated dictations to final expert annotations. The number below each step represents the final amount of output elements generated at this specific step. The output of one step is the input of the next step, except for step 5 for which we also include 5 randomly sampled examples from the seed dictations. At steps 6 & 7, the annotators also have access to the schema of step 3 along the synthetic dictation to produce its gold annotations.

Specifically, in the task data, we provide a hospital flowsheet schema that contains a list of all the rows (keys) in the flowsheet, the data type associated with each row (string, picklist, etc.), and, for rows associated with a picklist, a list of all valid values. It should be noted that the flowsheet schema can include hundreds to thousands of rows but it does not include any data specific to patients.

In addition, we provide a set of clinical transcripts where each transcript contains a realistic synthetic nurse dictation that mentions or describes a set of medical observations.

This task focuses on automatically populating the flowsheet by extracting all relevant medical information that are defined in the flowsheet and exist in each dictation.

3. Data Creation

In previous work (Corbeil et al., 2025b), we introduced SYNUR, a synthetic nursing dataset, which uses 80 proprietary fabricated seed dictations that were reviewed by practicing nurses for realism. Figure 2 summarizes the seven-stage pipeline, which alternates between the domain experts and the *GPT-4o-0806* language model.

1. **Observation mining**: The LLM iteratively extracted observation phrases (e.g., “dark yellow”) and linked them to clinical concepts (e.g., “urine colour”) from the 80 notes, yielding 547 unique observations.

2. **Concept consolidation**: We distilled these observations into 97 clinical concepts and assigned each a data type among boolean, integer, string, single-choice, or multi-choice.

3. **Ontology expansion**: Leveraging the LLM’s medical knowledge, we asked it to add relevant missing concepts with example sentences, expanding the ontology from 97 extracted concepts to 193 concepts, which an experienced nurse then validated and corrected.

4. **Scenario generation**: We prompted the LLM with the final concept set to craft coherent patient-case rationale and compatible observation combinations.

5. **Dictation synthesis**: Each scenario along with 5 randomly sampled seed dictations are passed to the LLM to generate a realistic nurse dictation that includes natural speech disfluencies and on-the-fly corrections.

6. **Train & development set annotation**: Expert nurse annotators verified the synthetic transcripts, and produced reference annotations using the ontology build at step 3 for a total of 223 dictations for training and development purposes that contain 3000 observations.

7. **Test set annotation**: We then produce a test set of 200 transcripts with an enhanced quality achieved by a disagreement-resolution step with the expert annotators.

For the MEDIQA-SYNUR 2026 shared task, we used the above mentioned 223 dictations as the training and development data—i.e., a 55:45 split, respectively. We also used the above mentioned 200 dictations as the test set that all teams were evaluated on.

Team	Affiliation	Paper	Code
Semantica Lab	University of Pennsylvania & Cedars Sinai Medical Center & University of Pittsburgh - USA	(Hwang et al., 2026)	1
MasonNLP	George Mason University - USA	(Karim and Özlem Uzune, 2026)	2
Smart_solutions	M42 - UAE	(Munjal, 2026)	3
HSE NLP	Higher School of Economics - Russia	(Valiev, 2026)	4
MKC	Seoul National University - South Korea	(Hwang and Kwak, 2026)	5
Gladiators	AbbVie - USA	(Satyanarayana et al., 2026)	6
MIDAS_SYNUR	MIDAS Lab, IIIT - India	(Sriram and Sahoo, 2026)	7
NSN	University of Pittsburgh & Washington University - USA	(Aryoyudanta et al., 2026)	8
Lakefront AI Ramblers	Loyola University Chicago - USA	(Saban et al., 2026)	9
AnotherOne	International Institute of Information Technology - India	(Thomas and Krishnamurthy, 2026)	10
LTRC-MEDICOM	International Institute of Information Technology - India	(Deepak et al., 2026)	11
SQUCS	Sultan Qaboos University - Oman	(JeebAllah et al., 2026)	12
LTRC-IIIT	International Institute of Information - India	(Vaish and Sharma, 2026)	13

1 https://github.com/syhwng/SemanticaLab-MediQA-SYNUR_submissions
2 https://github.com/AHMRezaul/MEDIQA-SYNUR-2026
3 https://github.com/PrateekMunjal/MEDIQA26_SYNUR_submission_Smart_solutions
4 https://github.com/Rebell-Leader/MEDIQA-SYNUR-HSE_NLP
5 https://github.com/KyominHwang/ClinicalNLP-2026-MKC
6 https://github.com/singhak-abbvie/MEDIQA-SYNUR-Gladiators
7 https://github.com/Swekrish303/MEDIQA-SYNUR-Team-MIDAS_SYNUR
8 https://github.com/yudanta/nsn-mediqa-synur
9 https://github.com/arsalanyaghoubi/Lakefront_AI_Ramblers-MEDIQA_SYNUR_2026
10 https://github.com/jr-john/mediqa-synur-2026
11 https://github.com/sushvinmarimuthu/ltrc-medicom-mediqa-synur-clinicalnlp-2026
12 https://github.com/RihamJeeballah/mediqa-synur-2026
13 https://github.com/av-dx/mediqa-synur-ltrc-iiit

Table 1: MEDIQA-SYNUR 2026: Participating teams, papers and codes.

4. Evaluation

4.1. Evaluation Metrics

For the task of extracting and normalizing clinical observations from nurse dictations and mapping them to a large flowsheet ontology of clinical concepts, we rely on the standard $F1$ metric for the evaluation of the extracted observations. In addition, we also present both precision (P) and recall (R) as secondary metrics for analysis purposes.

It should be noted that for each extracted multi-choice observation, we only mark it as true positive if it contains all the values that exist in the corresponding reference observation's list value. We provide a standardized evaluation script available on GitHub[2].

[2] https://github.com/abachaa/MEDIQA-SYNUR-2026

4.2. Code Verification

For additional validation, we required the submission of the code in addition to the models' outputs/runs, in order to verify the reproducibility of the reported runs. The participants shared their private codes with the organizers on GitHub following provided guidelines.

4.3. Baseline System

We built a GPT-4.1 based baseline system, with deterministic outputs (temperature=0). Our approach contains three sub-tasks (Corbeil et al., 2025b):

1. **Segmentation**: The segmentation step splits the streaming nurse transcript into medically coherent, continuous and non-overlapping segments via an LLM.

2. **RAG**: We minimize the size of the schema by filtering potential rows to the top N candidates given the current segment based on cosine similarity.

3. **Extraction**: Finally, we use the LLM to extract the relevant text from the transcript segment and canonicalize it based on the (now reduced) schema.

5. Official Results

5.1. Participating Teams

Thirteen finalist teams, out of 66 registered teams, submitted their code and runs for the MEDIQA-SYNUR 2026 shared task[3]. Table 1 presents the papers and the code for each team that participated. We limited the number of submitted runs to 15 runs per team.

5.2. Leaderboard Ranking

Table 2 presents the official results of the MEDIQA-SYNUR competition. Overall, the results show a relatively narrow performance gap among the top-ranked systems, with most approaches converging toward similar F1 scores despite substantial differences in model size and pipeline design.

Although excluded from the official ranking, the GPT-4.1 baseline (Corbeil et al., 2025b) achieved the highest overall performance with an F1 score of 0.837, outperforming the top ranked system by 2.3 absolute points. The baseline also achieved the highest precision and the second-highest recall, indicating that the proposed segmentation and RAG-based pipeline remains a very strong approach for the task.

Among the official submissions, the first-ranked system from *Semantica Lab* (Hwang et al., 2026) achieved an F1 score of 0.814 using a committee of frontier LLMs combined with an LLM-as-a-Judge verification strategy. This confirms the effectiveness of ensemble and verification-based pipelines, which appear frequently among the top-performing systems.

A notable observation is that systems ranked between second and fifth positions achieved very similar F1 scores (0.796–0.804), despite relying on different models and architectures. Many of these approaches were based on GPT-5 models or very large open-weight models such as Kimi-K2 MoE and Qwen3 235B, suggesting that performance differences in this range are driven more by pipeline design (e.g., retrieval, verification, or post-processing) than by the base model alone.

Different optimization strategies are also visible in the precision–recall trade-offs. For example, the *AnotherOne* system achieved the highest recall (0.835) but relatively low precision, while other top systems prioritized precision through verification or consensus mechanisms. This indicates that systems relying on verification stages tend to favor precision, whereas single-pass extraction systems often favor recall.

Another important result is the performance of smaller models. The *NSN* team successfully fine-tuned a Qwen3 2B model using SFT and GRPO, achieving an F1 score of 0.740 without complex pipeline components. This result is notable because it demonstrates that fine-tuned small language models can remain competitive with much larger models when properly trained for the task.

Finally, the lower-ranked systems (positions 11–13) relied primarily on prompting Llama-based models (8B and 70B), which resulted in lower overall performance. This suggests that prompting-only approaches without retrieval, verification, or fine-tuning are insufficient for this task compared to multi-stage pipelines or fine-tuned models.

Overall, the leaderboard highlights three main successful strategies: (i) multi-stage pipelines with retrieval and verification, (ii) ensembles or multi-model systems, and (iii) fine-tuning smaller models specifically for the task. These findings provide useful insights for future work on clinical observation extraction from nurse dictations.

5.3. Approaches

5.3.1. Semantica Lab

The *Semantica Lab* team (Hwang et al., 2026) introduced a 6-step mutli-stage pipeline that (i) normalize transcripts, (ii) generate a transcript-specific set of candidate concepts, (iii) filter candidates using an evidence gate or use an online router to select which domain experts to invoke for the transcript (iv) retain only extractions supported by cross-model agreement (v) send remaining ambiguous cases to an adjudicator. (vi) and canonicalize outputs and apply a consensus-based intersection of complementary runs.

5.3.2. MasonNLP

The *MasonNLP* team (Karim and Özlem Uzune, 2026) presented a modular retrieval-augmented extraction pipeline that contains the following components: (i) training-set exemplar retrieval, (ii) schema conditioned prompting (full schema or pruned candidate schema), (iii) deterministic schema-based postprocessing (iv) and a second-pass audit to refine outputs.

5.3.3. Smart_solutions

The *Smart_solutions* team (Munjal, 2026) adopted a multi-stage pipeline of agents. Specifically, they

[3] https://www.codabench.org/competitions/12113/

Rank	Team	Model(s)	Pipeline Components					Metrics		
			Desc.	Seg.	RAG	#LR	Ver.	P	R	F1
-	`Baseline`	GPT-4.1		✓	D	2	R	**0.843**	0.831	**0.837**
1	Semantica Lab	GPTs & Claudes				6	J	0.826	0.801	0.814
2	MasonNLP	GPT-5.2			D	2	V	0.786	*0.822*	*0.804*
3	Smart_solutions	Kimi-K2	✓		S	5		*0.805*	0.799	0.802
4	HSE NLP	Qwen-235B & GPT-5				3	J/X	0.781	0.819	0.800
5	MKC	GPT-5-mini & GPT-5.1	✓	✓	H	1		0.786	0.807	0.796
6	Gladiators	Claude Opus 4.5				1	R	0.731	*0.822*	0.774
7	MIDAS_SYNUR	GPT-5.2				1		0.779	0.767	0.773
8	NSN	Qwen3 2B $_{\text{SFT/GRPO}}$				1		0.701	0.785	0.740
9	Lakefront AI Ramblers	GPT-4o-mini		✓	H	2	V	0.768	0.662	0.711
10	AnotherOne	GPT-OSS 120B				4	R	0.607	**0.835**	0.703
11	LTRC-MEDICOM	Gemma2 & Llama3 8B				2	R	0.619	0.721	0.666
12	SQUCS	Llama3.3 70B				3	R	0.529	0.673	0.592
13	LTRC-IIIT	Llama3.1 8B				2		0.598	0.551	0.574

Table 2: Leaderboard results of MEDIQA-SYNUR (metrics: **P**recision, **R**ecall, **F1**). **Bold**, underlined and *italic* values are first, second and third scores per column, respectively. Several acronyms were used to differentiate steps across pipelines (excluding the mandatory extraction step): *Desc.* stands for description of concepts in the ontology; *Seg.* is for the transcript segmentation stage; *RAG* has the sparse (S), dense (D) and hybrid (H) systems; *#LR* stands for numbers of language-model roles; *Ver.* is for verification such as llm verification (V), llm-as-a-judge (J), rules (R), reparation (X), etc.

decomposed the main nursing task into 5 different subtasks: (i) observation extraction, (ii) ontology candidate retrieval, (iii) relevance scoring (iv) deterministic value assignment (v) post-processing and evaluation.

5.3.4. HSE NLP

The *HSE NLP* team (Valiev, 2026) introduced a three-stage pipeline that includes: (i) complementary generators that maximize candidate coverage (ii) a dedicated adjudicator that operates as a verifier rather than a generator, (iii) a token-efficient repair step that restores strict schema compliance.

5.3.5. MKC

The *MKC* team (Hwang and Kwak, 2026) proposed a RAG-based pipeline that first segment the input dictation into distinct units containing clinical facts. Afterwards, they constructed a memory bank for leveraging both a medical observation ontology and previously annotated observation tags from existing dictations. Finally, they integrated these dataset into a LLM generation process, in order to guide the model to produce more accurate outputs.

5.3.6. Gladiators

The *Gladiators* team (Satyanarayana et al., 2026) presented a hybrid approach that integrates large language models with domain-specific validation rules. Their system used a variety of comprehensive prompt engineering strategies and specialized filters and correction mechanisms. In addition, they included a supplementary rule-based component which includes a large set of regex patterns in order to captures high-confidence observations.

5.3.7. MIDAS_SYNUR

The *MIDAS_SYNUR* team (Sriram and Sahoo, 2026) showcased a prompting-based system that adopts a single-prompt, field-rich few-shot strategy, to jointly generate all schema fields in one structured output. Specifically, the few-shot demonstrations are curated and grouped by value type, in order to promote consistency across heterogeneous value distributions.

5.3.8. NSN

The *NSN* team (Aryoyudanta et al., 2026) presented a fine-tuned decoder only LLM for clinical observation extraction trained on the MEDIQA-SYNUR shared task dataset using an SFT followed

by GRPO-RL pipeline. In addition to the shared-task dataset, the team also incorporated an augmented CoT dataset during SFT, which improved model performance after GRPO-RL training.

5.3.9. Lakefront AI Ramblers

The *Lakefront AI Ramblers* team (Saban et al., 2026) introduced a hybrid schema retrieval with model/post-processing-level enhancements that includes the following steps: (i) segmentation of each transcript (ii) processing each segment by using RAG to select a small subset of schema concepts (iii) LLM-generated schema-constrained JSON extractions (iv) validation with optional post-processing steps (voting and LLM-based verification).

5.3.10. AnotherOne

The *AnotherOne* team (Thomas and Krishnamurthy, 2026) adopted a modular four-stage LLM pipeline that includes: (i) knowledge-enhanced concept detection using medical domain clustering (ii) evidence-grounded value extraction (iii) schema-constrained value normalization, and (iv) deterministic post-processing with fuzzy matching and unit pairing.

5.3.11. LTRC-Medicom

The *LTRC-Medicom* team (Deepak et al., 2026) presented a pipeline that consists of four distinct phases: (i) Semantic Clustering (ii) Systematic prompt engineering that includes a 9-step chain-of-thought system prompt (iii) Supervised Fine-Tuning using QLoRA (Quantized Low-Rank Adaptation) and (iv) transcript verification.

5.3.12. SQUCS

The *SQUCS* team (JeebAllah et al., 2026) presented a multi-agent LLM-based system that decomposes the extraction process into specialized agents responsible for (i) schema-guided extraction which includes an additional transcript segmentation mechanism (ii) rule-based validation (iii) and precision-oriented filtering verification. Finally, they also include a rule-based post-processing step informed by systematic error analysis on the development set.

5.3.13. LTRC-IIIT

The *LTRC-IIIT* team (Vaish and Sharma, 2026) introduced a pipeline that consists of two distinct stages: (i) a retrieval module designed to filter relevant flowsheet rows from the full set of schema rows (ii) a generation module that extracts values using a local instruction LLM.

6. Conclusion

The MEDIQA-SYNUR 2026 shared task was tackled by a wide variety of approaches from the participating teams. The runs submitted by the participating teams explored different retrieval-augmented generation, fine-tuning and prompting methods which we believe will provide new insights for future research directions in the task of observation extraction from nurse dictations.

The best performance was achieved from closed-weight model LLMs (e.g. GPT and Claude frontier models) but also teams that used very large open-weight model (e.g. Kimi-K2 MoE and Qwen3 235B) achieved similar results—in line with previous works (Dada et al., 2025; Corbeil et al., 2025a). A noticeable submission finetuned Qwen3 2B with SFT and GRPO achieving competitive results for such a small language model that was fine-tuned on a very small training set, paving the way to apply SLMs for the nurse observation extraction task.

Considering the diverse pipeline designs, we noticed a few successful strategies: adding descriptions to ontological concepts, segmenting the dictations, using RAG with dense embeddings, including several LLMs (i.e., agents or ensemble), and post-processing with LLM-as-a-Judge.

Incorporating all of these insights seems promising for the future of systems tackling the nurse observation extraction task. We hope that this shared task will encourage further efforts towards automatic observation extraction from nurse dictations to reduce the workload for medical professionals.

Limitations

The paper does not cover all possible methods and models for extracting clinical observations from nurse dictations. The SYNUR dataset is also limited in terms of size and types of clinical observations. Further experiments and evaluations are needed to validate the best performing methods on other datasets and scenarios.

Acknowledgments

We would like to thank Thomas Lin from Microsoft Health AI and the ClinicalNLP organizers for their support for the shared task. We also thank our annotation team, Molly Benton, Dominique Beck, Laurie Bonci, Ramona Brown, Nakisha Sanders, and Stefanie Barrera, for annotating the new SYNUR test set, as well as the participating teams who contributed to the success of the shared task through their interesting approaches and strong engagement.

7. Bibliographical References

Bayu Aryoyudanta, Maria Yuliana, Mikie Rachman, and I Made Agus Setiawan. 2026. Night shift nerds at mediqa-synur 2026: Pushing small large language model capability for clinical observation extraction and normalization from nurse dictation using rlvr. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Jean-Philippe Corbeil, Asma Ben Abacha, Jérôme Tremblay, Phillip Swazinna, Akila Jeeson Daniel, Miguel Del-Agua, and François Beaulieu. 2025a. Overview of the mediqa-oe 2025 shared task on medical order extraction from doctor-patient consultations. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 11–16.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025b. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 859–870, Suzhou (China). Association for Computational Linguistics.

Amin Dada, Osman Alperen Koraş, Marie Bauer, Jean-Philippe Corbeil, Amanda Butler Contreras, Constantin Marc Seibold, Kaleb E Smith, Jens Kleesiek, et al. 2025. Does biomedical training lead to better medical performance? In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 46–59.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Pasumarthy Deepak, Sushvin Marimuthu, and Parameswari Krishnamurthy. 2026. Ltrc-medicom at mediqa-synur 2026: Schema-guided clinical information extraction with hybrid clustering-sft-verification. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Bernal Jiménez Gutiérrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about gpt-3 in-context learning for biomedical ie? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512.

Kyomin Hwang and Nojun Kwak. 2026. Team mkc at mediqa-synur 2026: Retrieval-augmented generation based nurse observation extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Sy Hwang, Katherine S. Pitcher, Sue Hyon Kim, Yoonjae Lee, Hayoung K. Donelly, Harsh Bandhey, Andrew J. King, Karen O'Connor, Ryan J. Urbanowicz, and Danielle L. Mowery. 2026. Semantica lab at mediqa-synur 2026: Route, extract and verify – an llm-gated ensemble for parsing nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Riham JeebAllah, Adhari AlZaabi, and Abdulrahman Khalifa AAlAbdulsalam. 2026. Squcs at mediqa-synur 2026: A multi-agent open source llm system for nursing observation extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

A H M Rezaul Karim and Özlem Uzune. 2026. Masonnlp at mediqa-synur 2026: Retrieval-augmented large language models for schema-constrained clinical information extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Yanchi Liu, Wei Cheng, Haoyu Wang, Zhengzhang Chen, Mika Oishi, Takao Osaki, Katsushi Matsuda, et al. Improving open information extraction with large language models: A study on demonstration uncertainty. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Shazia Mitha, Jessica Schwartz, Mollie Hobensack, Kenrick Cato, Kyungmi Woo, Arlene Smaldone, and Maxim Topaz. 2023. Natural language processing of nursing notes: an integrative review. *CIN: Computers, Informatics, Nursing*, 41(6):377–384.

Prateek Munjal. 2026. Smart_solutions at mediqa-synur 2026: A multi-stage llm pipeline for nursing

observation extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

S. Panchal and P. Thakur. 2024. Harnessing the power of natural language processing in nursing services. *International Journal of Advanced Research*, 12(5):154–156.

Michael T. Saban, Arsalan Yaghoubi, Behnaz Eslami, Samie Tootooni, and Dmitriy Dligach. 2026. Lakefront ai ramblers at mediqa-synur 2026: Hybrid retrieval and llm verification for open-source schema-guided clinical information extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Siva Satyanarayana, Raju Pusapati, and Ankit Singh. 2026. Gladiator at mediqa-synur 2026: Contextual clinical extraction: Integrating foundation models with domain-specific validation rules. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Swetha Krishna Sriram and Akshitaa Sahoo. 2026. Midas_synur at mediqa-synur 2026: A prompting study for clinical observation extraction from nurse dictation transcriptions. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Jerrin John Thomas and Parameswari Krishnamurthy. 2026. Anotherone at mediqa-synur 2026: Detect, extract, normalize - knowledge-grounded llm pipeline for clinical observation extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Aashwin Vaish and Dipti Misra Sharma. 2026. Ltrciiit at mediqa-synur 2026: Benchmarking a fully local, training-free rag pipeline. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

Airat A. Valiev. 2026. Hse nlp team at mediqa-synur 2026: Consensus adjudication ensemble (ace): Balancing precision and recall for schema-bystander clinical extraction. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*, Mallorca Spain.

SemAnTICA Lab at MediQA-SYNUR 2026: Route, Extract, and Verify – An LLM-gated Ensemble for Parsing Nurse Dictations

Sy Hwang[1], **Katherine S. Pitcher**[1], **Sue Hyon Kim**[1], **Yoonjae Lee**[1], **Hayoung K. Donnelly**[1], **Harsh Bandhey**[2], **Andrew J. King**[3], **Karen O'Connor**[1], **Ryan J. Urbanowicz**[2], **Danielle L. Mowery**[1]

[1]University of Pennsylvania, [2]Cedars Sinai Medical Center, [3]University of Pittsburgh
Philadelphia, PA, USA; Los Angeles, CA, USA; Pittsburgh, PA, USA
{sy.hwang, hayoung.donelly, karoc, dlmowery}@pennmedicine.upenn.edu,
{kpitch, kshyon, yoonjael}@nursing.upenn.edu, {ryan.urbanowicz, harsh.bandhey}@cshs.org,
andrew.king@pitt.edu

Abstract

We describe the **Sem**antic **An**alysis of **T**ext to **I**nform **C**linical **A**ction (SemAnTICA) Lab's system for the MediQA-SYNUR 2026 shared task on extracting structured clinical observations from nurse dictation transcripts. The task requires mapping observations from disfluent conversational text to a large, fixed ontology and producing strictly normalized outputs, where small amounts of concept over-selection severely degrade micro F1 score. Our approach evolved from a full-schema in-context baseline to a pipeline that explicitly separates concept selection from value extraction. We first preprocess transcripts, then generate transcript-specific concept candidates using hybrid sparse–dense retrieval. The candidates are then pruned with an evidence-based filter. For extraction, we adopt a system-level mixture-of-experts design with an online Large Language Model (LLM) router that selects a subset of domain-specialized experts per transcript. Each expert operates over a constrained schema partition to reduce spurious predictions. We enhance robustness with agreement-gated ensembling and targeted adjudication for ambiguous cases. Finally, we intersect complementary high-recall and high-precision runs to produce the best submission. Our system ranked first on the official test leaderboard with $F_1 = 0.814$, $P = 0.826$, $R = 0.801$.

Keywords: information extraction, ontology alignment, nursing documentation, large language models, verification

1. Introduction

In clinical care, patient documentation is both mission critical and burdensome, and nursing workflows are particularly documentation-intensive. To reduce the burden of documentation for nurses, voice recognition (VR) systems have been implemented. While the use of VR has been shown to decrease the time spent on documentation, nurses mainly document via the keyboard in electronic health record (EHR) flowsheets (Mayer et al., 2021). For VR to be more widely adopted, methods must be implemented to extract the relevant clinical data from nurse dictation transcriptions and translate this information into a format compatible with EHR workflow data. MediQA-SYNUR targets this challenge by converting informal nurse dictation transcripts into structured flowsheet-like observations. Two factors drive the difficulty of the automation of this task: (1) spoken language informality, noise and imprecision (Galatzan and Carrington, 2018) and (2) the large, heterogeneous output space. Systems must identify the relevant text, select the correct concept from a sizable ontology and then produce an appropriate value type. Depending on the concept, the value may come from an enumeration, take the form of a numeric measurement with units, or be unconstrained free text.

When attempting to extract and canonicalize outputs to an ontology across a large number and range of clinical concepts, several challenges emerge. At scale, concept selection is the dominant factor affecting performance, as even modest over-selection produces many false positives under strict matching. Free-text `STRING` values further increase difficulty: without a fixed value set, models must determine both extraction boundaries and surface form, while exact-match evaluation penalizes any paraphrasing. Additionally, we observed that some `STRING` references can vary stylistically across otherwise similar evidence, which compounds mismatch risk. These properties motivate a design that treats concept selection as a retrieval-and-verification problem and uses explicit uncertainty control during extraction.

Our approach follows three design principles: separate candidate generation from value extraction via retrieval and evidence checks; use online routing to invoke only relevant domain experts; and control variance with agreement-gated ensembling plus targeted adjudication for ambiguous cases. We describe the evolution of this system and the configurations used in our submissions for the shared task.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 27–37
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Related Work

Schema-guided information extraction has long been a central strategy for normalizing clinical text into structured representations, particularly when downstream use requires consistent concept identifiers and typed values. MediQA-SYNUR sits squarely in this tradition, with the added challenge that the input is spoken dictation rather than edited clinical notes. The shared task is described in the official overview paper (Michalopoulos et al., 2026), and it builds on the SYNUR dataset introduced by Corbeil et al. (2025), which frames structured observation extraction from nurse dictations as a high impact but underexplored clinical NLP problem due to limited public data availability. As part of the broader MediQA series, the task also inherits a shared-task lineage emphasizing standardized evaluation and comparative system analysis across participants (Ben Abacha et al., 2019, 2021).

Prior work at the intersection of nursing workflows and clinical NLP has explored both speech-derived text and traditional nursing documentation. Song et al. (2022) examine NLP over auto-generated transcripts of patient–nurse communication in home health care, showing that conversational, Automatic Speech Recognition (ASR)-derived text can support downstream risk identification while highlighting noise sources that complicate reliable extraction. King et al. (2023) similarly integrate automatic transcription with NLP in a voice-based digital assistant designed to support Intensive Care Unit (ICU) rounds, illustrating the feasibility of speech-driven clinical decision support and the importance of robust normalization and evidence-grounded prompting. Complementing these speech-oriented efforts, Mitha et al. (2023) review a broader landscape of NLP applied to nursing notes, synthesizing common tasks and methodological patterns in nursing documentation analytics. Together, these studies motivate MediQA-SYNUR's focus on nurse-centered language and workflow realism, while underscoring a gap addressed by the shared task: ontology scale, schema-guided normalization of nursing observations from dictation transcripts under strict evaluation.

Recent clinical NLP systems increasingly employ large language models (LLMs) for extraction and normalization, motivated by their strong instruction following and few-shot generalization (Rodrigues and Teixeira Lopes, 2025). However, ontology scale constraints can shift the dominant failure mode away from value interpretation and toward concept over-selection. When many plausible schema elements are presented, models may default to exhaustive slot filling, producing spurious outputs that degrade precision under strict matching. This issue mirrors observations in other structured medical information extraction pipelines where specification, normalization conventions, and schema constraints strongly shape end performance, and it motivates architectures that separate candidate generation from value decoding.

Retrieval-augmented generation is often framed as a method for injecting external knowledge into generation (Lewis et al., 2020), yet retrieval can also play a purely structural role by selecting which schema elements should be considered for a given input. In schema-guided extraction, this retrieval-as-constraint view is particularly natural: candidate concept selection becomes an information retrieval problem over schema descriptors, synonyms, and notes, and subsequent extraction is performed only over a reduced candidate set.

Robustness methods for LLM-based systems frequently rely on ensembling and adjudication. Multi-model ensembles can reduce variance and provide a practical uncertainty signal through agreement, while selective verification concentrates compute on ambiguous cases. In parallel, "LLM-as-a-judge" approaches have grown into a widely studied paradigm for assessment and decision-making (Li et al., 2025).

Agentic and tool-using LLM systems emphasize input-conditional orchestration, where a controller selects actions or tools based on the current instance. Yao et al. (2023) formalize an interleaving of reasoning and acting, while Schick et al. (2023) demonstrate learning to decide when to call external tools. Related modular architectures such as the Modular Reasoning, Knowledge and Language (MRKL) system describes routing queries to specialized modules, highlighting the value of a router that dispatches work to domain-appropriate components (Karpas et al., 2022). In our setting, the closest analogue is an online LLM router that selects which domain-specialized extraction experts to invoke per transcript. We describe this as a system-level mixture-of-experts (MoE) design with LLM-based gating, conceptually related to conditional computation and routing ideas popularized in sparse MoE models (Lepikhin et al., 2021; Fedus et al., 2022), while differing in that our experts are separate LLM calls rather than trainable subnetworks in a single model.

3. Data and Task

The MediQA-SYNUR task provides annotated synthetic nurse dictation transcripts and a fixed ontology of observation concepts. Each output record links a clinical observation mention in the transcript to a schema concept and a typed value. The schema spans multiple clinical domains and includes four value types: categorical `SIN-`

GLE_SELECT (choose exactly one option), categorical MULTI_SELECT (choose one or more options), NUMERIC measurements, and free-text STRING.

Figure 1 summarizes two corpus-level properties that shape our system design. (a) The distribution of annotated observations by value type shows that categorical and numeric fields constitute the majority of labels, while unconstrained STRING fields are a small fraction. (b) The distribution of the number of annotated observations per transcript in the training and development splits shows a similar central tendency (≈ 13 observations per sample). Together, these trends indicate that the effective output space per transcript is small relative to the full ontology, and that most supervision is concentrated in non-STRING types. This motivates curation of schema concepts specific to each transcript, and conservative handling of unconstrained free-text values under exact-match scoring.

The transcripts are synthetic but resemble clinical speech, containing hedges and fillers, abbreviated phrasing and implicit references. Because the evaluation requires exact matching of normalized outputs, the task rewards conservative, evidence-backed extraction. In practice, this means that a system must balance recall against the risk of over-selecting concepts that are only weakly implied.

4. Methods

4.1. Overview

Our system is a multi-stage pipeline that gradually narrows uncertainty. We (1) normalize and clean transcripts, then (2) generate a transcript-specific set of candidate concepts via retrieval. Next, we either (3) filter candidates using an evidence gate that removes concepts without direct textual support, or (4) use an online router to select which domain experts to invoke for the transcript. Each expert extracts observations from a constrained subset of the schema. When we run multiple models within an expert, (5) we retain only extractions supported by cross-model agreement and send remaining ambiguous cases to an adjudicator. Finally, (6) we canonicalize outputs and apply a consensus-based intersection of complementary runs.

Throughout, we view the system as taking on two distinct decisions: identifying which concepts are present in the transcript, and then extracting their values. In our error analysis, misses were driven primarily by the first step, concept selection, rather than by value extraction.

4.2. Transcript pre-processing

We pre-process transcripts to reduce superficial variability that confuses both retrieval and generation. Specifically, we repair encoding artifacts when present (e.g. "\u00e2\u20ac\u201d"), removing speaker tags (e.g. "[Clinician]") and conversational markup (e.g. "Okay, let's see here"), and lightly filtering disfluencies that do not carry clinical meaning (e.g. "uh" or "um"). This pre-processing improves lexical and semantic matching between transcript phrases and schema descriptors.

4.3. Schema enrichment

Early error analyses showed that surface-form mismatch was a major source of both missed concepts and unstable retrieval. To improve coverage, we enriched clinical concepts from the schema with concise synonym strings and concept-specific normalization guidance that defines how values should be represented. This information is injected at extraction time as a compact constraint, reducing variation in generated values. For example, when a transcript indicates that an assessment was performed, the notes can specify a preferred normalized value rather than allowing the model to invent phrasing.

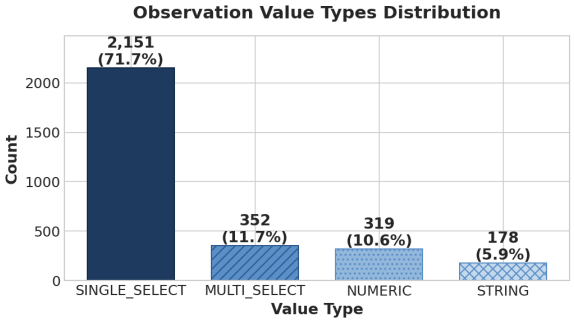
4.4. Candidate concept generation via hybrid retrieval

Presenting all schema concepts to a single extractor encourages over-selection. Instead, we generate a transcript-specific candidate set using retrieval over schema descriptors (concept name, synonyms, and notes). We use sparse lexical matching via term frequency-inverse document frequency (TF-IDF) to capture exact phrase overlap and dense embeddings to capture semantic similarity via cosine distance. A hybrid fusion strategy balances the two, allowing synonymy to be recovered without losing the precision of lexical anchors.

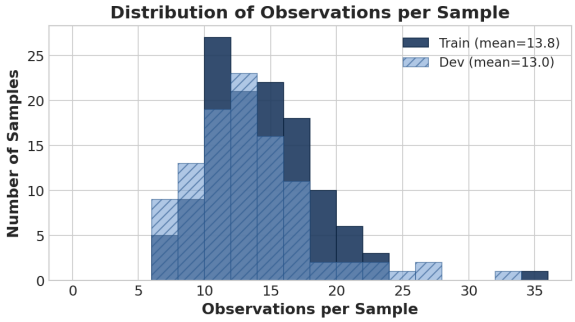
This stage is best understood as candidate generation in an information retrieval pipeline. It is intentionally permissive, but it should shrink the effective output space from the full ontology to something closer to the expected number of observations per transcript.

4.5. Evidence-based candidate filtering

Hybrid retrieval improves recall but can still propose candidates that are merely thematically related to the transcript. To reduce false positives, we apply a lightweight evidence check. For each candidate concept, a model must point to explicit supporting text or make a tightly constrained inference that is defensible from the transcript. Candidates that cannot be justified are removed. This step shifts the system from semantic association toward evidence-grounded selection, which is essential under strict evaluation.



(a) Distribution of value types.



(b) Distribution of the number of observations per sample.

Figure 1: Data characteristics of MEDIQA-SYNUR training and development sets: (a) value-type distribution; (b) observations per transcript.

4.6. Few-shot example retrieval

Many remaining errors reflect normalization conventions, including categorical choice, measurement formatting, and output structure. To stabilize these choices, we retrieve a small set of similar training transcripts and include their labeled outputs as demonstrations. Few-shot examples are selected using the same representations used for schema retrieval, aligning the model's context with the transcript's phrasing.

4.7. Value extraction with online LLM routing and domain experts

4.7.1. Reduced-schema global extraction

In the reduced-schema variant, we expose a single extractor, or a committee of extractor models, to the transcript's candidate concepts distilled from the retrieval process, rather than the full ontology. This proves effective when candidate reduction is sharp, but can still leave the model facing a large and heterogeneous schema when many concepts remain plausible. We therefore introduce online routing as a complementary strategy for managing schema overload. Instead of one global pass over a broad reduced schema, we select a small set of domain-partitioned experts per transcript to limit full ontology exposure.

4.7.2. Domain-specialized experts with constrained output spaces

To mitigate prompt overload and curb concept over-selection, we decompose the full ontology into domain-specialized extraction experts (e.g. cardiovascular or mobility), each responsible for a disjoint subset of schema concepts (approximately ≤ 20 concepts per expert). Each expert is implemented as an LLM call with a domain-specific prompt containing only its assigned concepts, which constrains the output space and reduces the likelihood of spurious extractions. This design functions as a system-level mixture-of-experts, where "experts" are domain-specialized LLM extractors and constraint is enforced by the schema partition.

Figure 2 illustrates our system-level mixture-of-experts design. Given a nursing dictation transcript, an LLM router selects a subset of clinical domains (optionally subject to a routing budget K). Each selected domain expert operates over a disjoint schema partition and extracts candidate observations constrained to its domain-specific schema and normalization rules. For robustness, each domain expert can be instantiated as a small committee of heterogeneous LLMs; extracted observations are aggregated using an agreement gate over $(\texttt{concept_id}, \texttt{value})$ pairs (accept if at least 2/3 models agree). Accepted observations are merged across domains to form the final structured output.

We define domains aligned with clinical practice patterns (e.g., vital signs, respiratory, neuro exam, GI, GU, skin/wound, mobility/safety, behavioral risk, pain, and general patient information). Domain decomposition substantially reduces per-call schema size and improves precision by narrowing attention to semantically related concepts.

4.7.3. Online LLM routing

Rather than invoking all experts for every transcript, we use an online LLM router to select which domain-specialized experts to run for each input. The router reads the normalized transcript and returns a subset of relevant domains, optionally with confidence scores or an expert budget. Routing is input-conditional: a transcript centered on shortness of breath and oxygen therapy may activate respiratory and cardiopulmonary experts, whereas a wound-care update may activate skin/wound and mobility/safety experts. This reduces unnecessary expert calls and directly targets the main failure mode observed in early submissions: inflated numbers of predicted observations caused by ontology-

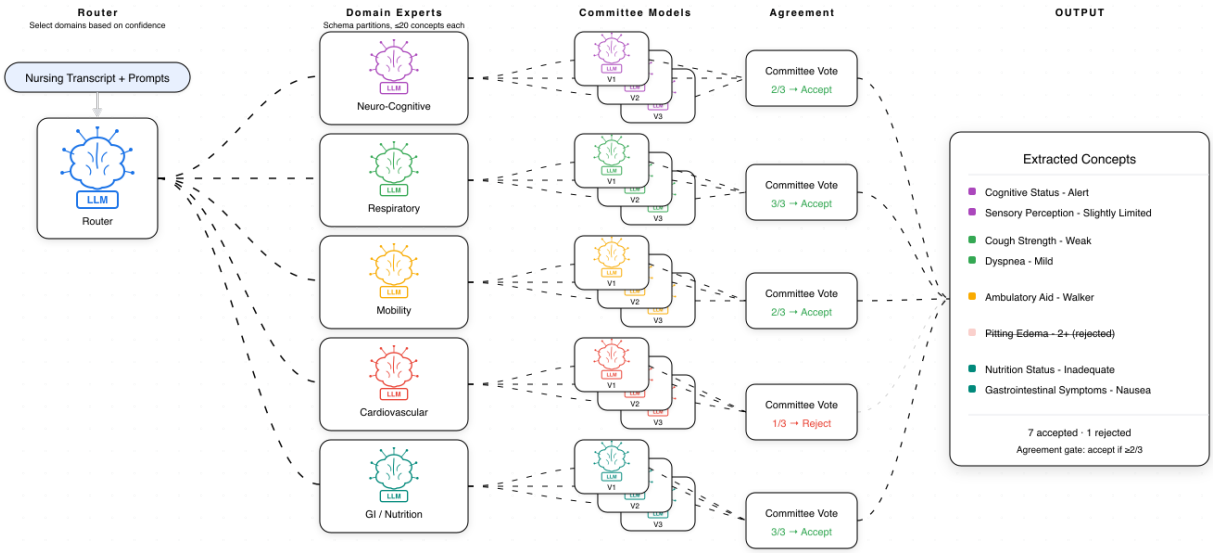


Figure 2: Ontology partitioning into domain-specialized experts (Section 4.7.2).

scale over-selection.

In implementation, the router is prompted to be conservative. It is instructed to select only domains supported by explicit transcript evidence and to avoid precautionary or "just in case" routing. When router confidence is low, or when a transcript appears genuinely multi-domain, we allow a larger expert budget and rely on downstream agreement gating and validation to control false positives. The effective subset of concept space considered by the system is therefore input-dependent rather than fixed.

This routing step gives the system a mixture-of-experts style architecture at the system level: the router acts as an LLM-based gating mechanism that determines which specialized extractors are applied to a given transcript. We use "mixture-of-experts" in this operational sense only; we do not train an end-to-end learned gating network.

4.8. Agreement-gated ensembles and targeted adjudication

Within each routed domain expert, we optionally run a committee of heterogeneous LLMs and aggregate extractions using agreement-gated voting over $(\texttt{concept_id}, \texttt{value})$ pairs. Items failing the agreement threshold, or items with small margins between competing outputs, are routed to targeted adjudication: a separate verifier model re-checks transcript evidence and accepts or rejects disputed items. Because adjudication is triggered only for ambiguous cases, it provides a favorable cost–accuracy trade-off relative to always-on verification. Importantly, this design complements online routing: routing limits which experts run, and agreement/adjudication limits which extractions survive.

4.9. `STRING` handling and competition-time decision

`STRING` concepts behaved differently from categorical and numeric concepts. Without enumerated value sets, the model must determine both the boundary of the extracted text and its canonical phrasing. Under exact-match evaluation, even semantically correct paraphrases score as errors. In addition, clinician review of training and development set reference standards suggested that a significant portion of `STRING` references can be inconsistent across similar transcript evidence. To avoid optimizing against noisy supervision, and given the disproportionate false positive risk introduced by unconstrained values, we adopted a pragmatic policy to exclude `STRING` concepts from output in configurations where doing so improved overall micro F1. This decision is not a claim that `STRING` extraction is unimportant; rather, it reflects the difficulty of unconstrained value generation under strict scoring in noisy transcripts.

4.10. Validation and canonicalization

We apply deterministic post-processing to enforce schema constraints, ensure type consistency, normalize casing and units where applicable, and remove invalid or malformed outputs. This stage also protects against model "creativity" by rejecting outputs that do not match the schema's expected structure.

4.11. Late fusion rule for the final run

In the final phase, we treated system combination as a decision-level fusion rule rather than a run-specific heuristic. We maintained two complementary extractors: a high-recall configuration that

favors broader candidate generation and routing, and a high-precision configuration that uses conservative routing with agreement gating. Because ontology scale evaluation heavily penalizes spurious concepts, we used cross-system agreement as a confidence signal at the concept level. Concretely, we accepted a concept only when it was predicted by both systems, and we selected the corresponding value from the configuration with the lower value-format error rate (the high-precision system). This rule can be viewed as a special case of thresholded multi-system voting and was chosen based on development-set calibration to control false positives without materially reducing recall.

4.12. Implementation Details

This subsection summarizes the principal implementation choices needed to understand and approximately reproduce the system configurations evaluated in this study. Unless stated otherwise, hyperparameters were tuned on the training/development splits and fixed for inference.

Schema candidate retrieval. Each schema concept is serialized as a text string comprising the concept name and auxiliary descriptors (e.g., synonyms, concise notes). We employ a hybrid retrieval strategy: sparse retrieval utilizes a TF–IDF vectorizer over unigrams and bigrams with L_2 normalization, while dense retrieval computes cosine similarity between transcript and concept embeddings using OpenAI's `text-embedding-3-small`. Given the dataset scale, we employ brute-force cosine scoring. The sparse and dense signals are combined via linear score fusion after min–max normalization:

$$s(c \mid x) = \alpha \cdot \tilde{s}_{\text{tfidf}}(c \mid x) + (1-\alpha) \cdot \tilde{s}_{\text{emb}}(c \mid x), \quad (1)$$

where $\alpha \in [0, 1]$ is a run-specific hyperparameter (see Table 2). We retain the top-k candidates (typically 40–60) for downstream processing.

Few-shot selection. Demonstrations are retrieved from the training set using the same hybrid TF–IDF/embedding strategy. We select k_{shot} examples (typically 3–10) and format them as transcript–label pairs using the official output schema to encourage structural consistency.

Evidence gating. To mitigate false positives from permissive retrieval, we introduce an evidence gate. This module verifies candidate concepts against the transcript, rejecting those lacking explicit support. A candidate is retained only if the model identifies a supporting span and the value type aligns with the evidence.

Online routing and expert invocation. Domain-specialized experts are implemented as distinct extraction prompts, each restricted to a disjoint schema partition (≈ 20 concepts per domain). A router model (GPT-5) processes the transcript and domain descriptions to output: (i) a subset of relevant domains, and (ii) a ranked list of concept candidates with confidence scores. We enforce a global budget by selecting the top-K candidates based on router confidence. These selections determine which expert prompts are invoked and constrain the schema exposed to each expert.

Models and decoding. For extraction, we utilize temperature $T = 0.2$. For reasoning-capable models, we set `reasoning_effort="high"`. All components (router, expert, and judge) are constrained to emit a fixed JSON schema of (*concept_id*, *value_type*, *value*) tuples to ensure deterministic parsing. Generation limits are calibrated to cover expected observation counts while penalizing verbosity. Unless otherwise specified in Table 2, the same family of LLMs used for extraction was also used for evidence-gated candidate verification, while GPT-5 served as the router and, in most committee configurations, the adjudication model.

Committee aggregation and adjudication. When operating as a committee, outputs are aggregated by voting. A concept-value pair (c, v) is accepted if its support fraction exceeds a threshold τ_{agree}:

$$\frac{1}{n}\sum_{i=1}^{n}[(c, v) \in O_i] \geq \tau_{\text{agree}}, \quad (2)$$

where O_i is the output set from voter i. Conflicts where competing values for a single concept have a support margin difference below δ are flagged and optionally routed to a judge model for evidence-based resolution.

Post-processing and canonicalization. We filter invalid concept identifiers and malformed JSON. Categorical values are mapped to canonical schema labels, while numeric fields undergo normalization (e.g., unit standardization, decimal formatting). Final outputs are deduplicated and serialized into the submission format.

Decision-level fusion. The final submission utilizes decision-level fusion across two independently configured systems. A concept is extracted only if predicted by both systems (cross-system agreement). The value is selected from the system demonstrating lower value-format error on the development set, serving as a conservative consensus filter.

Prompt and schema refinement. We iteratively refined prompts using the development set, guided by structured review from three clinically trained nurses on our team. Review focused on recurring failure modes such as prediction consistency, value normalization, and missed mentions. In parallel, the nurse reviewers expanded the schema with additional surface forms such as synonyms and common abbreviations for high frequency concepts to improve candidate recall during retrieval and reduce brittle exact-match errors. All refinements were conducted using only the shared task development data. Clinician input was limited to qualitative review and schema expansion. No additional labeled examples beyond the released development annotations were created. We did not use and did not have access to the test set for prompt or schema updates.

5. Results

5.1. Official leaderboard performance

Our best run (Run 526481 on Codabench) achieved $F_1 = 0.814$, $P = 0.826$, $R = 0.801$, ranking first on the official test leaderboard. A comparison of the ablation and fusion variants we evaluated is provided in Appendix A.

5.2. What improved performance late in development

The strongest improvements came not from better prompting in isolation but from controlling the number of plausible concepts a model is asked to consider at once. Our initial test submissions exhibited high recall but weaker precision, consistent with inflated predicted observation counts. Introducing online expert routing reduced irrelevant concept exposure, and domain specialization made it easier for extractors to stay faithful to transcript evidence. Agreement gating further dampened variance, while targeted adjudication prevented ambiguous edge cases from dominating errors. Results for each configuration from our test submissions and their corresponding performance on the development set are provided in Table 1.

As a shared task system, our goal was to assemble a reliable high-precision extraction pipeline rather than to isolate each module under a fully factorial experimental design. The reported comparisons therefore support the utility of the overall architecture and several major design choices, but do not fully disentangle all higher-order interactions among routing, agreement gating, adjudication, and late fusion.

5.3. Qualitative behavior of the final system

The final system exhibited a more conservative extraction profile: it was less likely to output observations that were merely plausible in context, and more likely to output observations with explicit textual support. This behavior is reflected in the mean number of predicted observations per transcript, $\bar{x}_{\text{obs}}$, which was released to participants after the submission phase closed. The gold test set label density had $\bar{x}_{\text{obs}} = 13.61$, closely matching our final system's predicted $\bar{x}_{\text{obs}}$. Concept-level intersection filtering provided an additional safety layer by removing low-confidence, single-run concepts while preserving the core set of clinically salient extractions.

6. Discussion

6.1. Ontology scale makes concept selection the bottleneck

A central lesson of this shared task is that ontology scale changes the nature of extraction. When the number of candidate concepts is large relative to the number of true observations per transcript, false positives become the dominant failure mode. Even if a model can accurately extract values conditional on the correct concept, overall performance degrades if it cannot reliably decide which concepts are actually present. Candidate generation, evidence filtering, and routing are therefore not ancillary utilities; they are the primary mechanisms that turn a general-purpose generator into a disciplined, ontology scale extractor.

6.2. Online routing is a practical form of structured inference

Online routing offers a middle ground between monolithic prompting and full learned MoE. It uses an LLM's semantic understanding to decide where to spend attention and compute, but it preserves strict output constraints by delegating extraction to experts with narrow schema partitions. In practice, routing behaves like structured inference: it encourages the system to commit to a small set of hypotheses and to ignore the rest.

6.3. Agreement as uncertainty, adjudication as targeted verification

Ensembling is often used for robustness, but its most useful role here was as a proxy for uncertainty. Disagreement reliably marked concepts that were weakly evidenced, ambiguously phrased, or prone to normalization variation. By adjudicating only those cases, we achieved much of the benefit

Table 1: Development and test set performance across system configurations. Rows are sorted by test F1 (descending).

	Development Set			Test Set			
run	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1** ↓	$\bar{x}_{\text{obs}}$
A	0.860	0.870	0.865	**0.826**	0.801	**0.814**	13.26
B	**0.868**	0.863	*0.865*	0.802	0.813	*0.807*	13.86
C	*0.862*	*0.870*	**0.866**	0.800	0.813	0.807	13.89
D	0.842	0.867	0.854	0.767	0.845	0.804	15.06
E	0.856	0.843	0.850	*0.802*	0.778	0.790	13.26
F	0.823	**0.882**	0.851	0.707	**0.853**	0.773	16.51
G	0.843	0.819	0.831	0.738	0.806	0.770	13.89
H	0.821	0.833	0.827	0.654	*0.848*	0.738	17.74
I	0.753	0.797	0.774	0.719	0.630	0.671	11.98
J	0.772	0.724	0.747	0.706	0.637	0.670	12.34

of verification without the cost of verifying every extraction.

6.4. `STRING` remains an open challenge under strict scoring

The difficulty of free-text extraction under exact-match evaluation is not unique to this task, but the combination of spoken dictation and inconsistent normalization can make it especially severe. Our decision to exclude `STRING` in some configurations reflects a reality: unconstrained outputs create a large surface for mismatches, and a small number of systematic `STRING` false positives can dominate micro F1. Future shared tasks may benefit from evaluation protocols that reward semantic correctness or span fidelity for free-text fields.

6.5. Annotation protocol shift across splits

During error analysis, our clinical reviewers noted apparent inconsistencies in the normalization of some observations in the development set reference standard, most prominently for unconstrained `STRING` concepts where similar transcript evidence could correspond to different reference values. In follow-up communication, the organizers clarified that the training and development splits each underwent a single round of annotation, with multiple trained annotators labeling different portions of the data, whereas the test split received an additional round of annotation by different annotators followed by reconciliation to resolve disagreements and improve consistency. This design is a reasonable allocation of annotation effort for shared tasks and likely increases the reliability of the held-out evaluation. At the same time, differing annotation workflows across splits can introduce an annotation protocol shift, where systems tuned on noisier or more heterogeneous training and development labels may generalize differently under a more consistent test reference standard. This effect is especially salient for free-text normalization, where stylistic variation and boundary ambiguity amplify disagreement. Future iterations could strengthen interpretability by applying comparable adjudication procedures across splits when feasible, or by reporting split-specific inter-annotator agreement and reconciliation policies to clarify expected test set behavior.

7. Conclusion

We presented SemAnTICA Lab's MediQA-SYNUR 2026 submission, a pipeline that treats ontology scale extraction as a problem of disciplined concept selection and controlled generation. By separating candidate generation from value extraction, adding evidence-based filtering, and introducing an online LLM router to gate domain-specialized experts, we improved precision without sacrificing recall. Agreement-gated ensembling and targeted adjudication further stabilized outputs, and a consensus-based intersection of complementary runs produced our best leaderboard result. More broadly, the shared task underscores that schema scale and value type heterogeneity remain central obstacles for dictation to structure extraction. Our results suggest that these challenges are best addressed not by a single monolithic model call, but by disciplined selection and validation of candidates through retrieval, expert routing, and consensus systems. We expect future work to build on this principle by improving evidence calibration and value normalization while retaining the same system-level guardrails.

8. Limitations

Our system improves precision through staged control, but this comes with important tradeoffs. Retrieval, routing, committee inference, and optional

adjudication increase latency and cost relative to simpler baselines, and our best configurations rely on multiple external LLMs whose availability, pricing, and institutional acceptability may change over time. Because performance emerges from the interaction of several modules, the contribution of any single component is only partially isolated by our ablations. In addition, the system was tuned for a benchmark with synthetic nurse dictations and strict exact-match scoring, so its behavior may differ in real clinical documentation settings, especially for unconstrained free-text fields. Finally, our decision to exclude STRING in some configurations reflects benchmark-specific optimization rather than a general solution to free-text extraction.

9. Ethics Statement

The task is motivated by reducing documentation burden, but automated extraction systems can omit, distort, or overstate clinically relevant information. Any real deployment should include human oversight, careful auditing, and calibration for high risk concepts. Our work reports aggregate performance and common failure patterns and does not attempt to infer sensitive attributes beyond the task's intended outputs from this synthetic dataset.

10. Bibliographical References

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.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy. Association for Computational Linguistics.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Suzhou (China). Association for Computational Linguistics.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Benjamin J. Galatzan and Jane M. Carrington. 2018. Exploring the state of the science of the nursing hand-off communication. *Computers, Informatics, Nursing: CIN*, 36(10):484–493.

Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, Dor Muhlgay, Noam Rozen, Erez Schwartz, Gal Shachaf, Shai Shalev-Shwartz, Amnon Shashua, and Moshe Tenenholtz. 2022. MRKL systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *CoRR*, abs/2205.00445.

Andrew J. King, Derek C. Angus, Gregory F. Cooper, Danielle L. Mowery, Jennifer B. Seaman, Kelly M. Potter, Leigh A. Bukowski, Ali Al-Khafaji, Scott R. Gunn, and Jeremy M. Kahn. 2023. A voice-based digital assistant for intelligent prompting of evidence-based practices during icu rounds. *Journal of Biomedical Informatics*, 146:104483.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

LeAnn Mayer, Dongjuan Xu, Nancy Edwards, and Gordon Bokhart. 2021. A comparison of voice recognition program and traditional keyboard charting for nurse documentation. *Computers, Informatics, Nursing: CIN*, 40(2):90–94.

George Michalopoulos, Jean-Philippe Corbeil, Cari Bader, Nate Bodenstab, and Asma Ben Abacha. 2026. Overview of the MEDIQA-SYNUR 2026 shared task on observation extraction from nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

Shazia Mitha, Jessica Schwartz, Mollie Hobensack, Kenrick Cato, Kyungmi Woo, Arlene Smaldone, and Maxim Topaz. 2023. Natural language processing of nursing notes: An integrative review. *CIN: Computers, Informatics, Nursing*, 41(6):377–384.

Tiago Rodrigues and Carla Teixeira Lopes. 2025. Harnessing large language models for clinical information extraction: A systematic literature review. *ACM Trans. Comput. Healthcare*, 6(4).

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Proceedings of the 37th Conference on Neural Information Processing Systems*.

Jiyoun Song, Maryam Zolnoori, Danielle Scharp, Sasha Vergez, Margaret V. McDonald, Sridevi Sridharan, Zoran Kostic, and Maxim Topaz. 2022. Is auto-generated transcript of patient-nurse communication ready to use for identifying the risk for hospitalizations or emergency department visits in home health care? a natural language processing pilot study. In *AMIA Annual Symposium Proceedings*, pages 992–1001. Published in 2023 as AMIA Annu Symp Proc. eCollection 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

A. Experiment Configurations

Table 2 details the full configuration for each experimental run.

Table 2: Experiment Configurations

run	fusion	full	STR	part	tfidf_s	emb_s	tfidf_f	emb_f	top_k	fs_k	agree	margin	router	committee	judge
A	Run B+D	No	No	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
B	None	No	No	Yes	NA	NA	0.3	0.7	60	5	0.6	0.15	NA	• gpt-5 • claude-sonnet-4.5 • gpt-o3 • gpt-oss-120b • gpt-4.1	gpt-5
C	Run B+F	No	No	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D	None	No	No	No	0.3	0.7	0.3	0.7	50	5	0.6	0.1	30	• gpt-5 • claude-sonnet-4.5 • gpt-4.1	gpt-5
E	None	No	No	Yes	NA	NA	0.3	0.7	40	6	0.7	0.15	NA	• gpt-5 • claude-sonnet-4.5 • gemma-3-12b • gpt-o3 • gpt-oss-20b • gpt-4.1	NA
F	None	No	No	No	0.4	0.6	0.4	0.6	50	5	0.6	0.2	30	• gpt-5 • gpt-oss-120b • claude-sonnet-4.5	gpt-5
G	None	No	No	No	0.5	0.5	0.5	0.5	60	10	0.7	0.3	60	• gpt-5 • gemma-3-12b • gpt-o3 • gpt-oss-20b • gpt-4.1	gpt-5
H	None	Yes	No	No	NA	NA	NA	NA	NA	0	NA	NA	NA	• gpt-5	NA
I	None	No	Yes	No	0	1	0	1	50	3	NA	NA	30	• gpt-5	NA
J	None	No	Yes	No	1	0	1	0	50	3	NA	NA	20	• gpt-5	NA

Column definitions

Run-level: `run` (run ID); `fusion` (decision-level fusion rule).

Prompt / eligibility: `full` (full ontology vs. candidate subset); `STR` (STRING concepts inclusion); `part` (schema partitioning with domain experts).

Schema retrieval: `tfidf_s`/`emb_s` (sparse vs. dense weights); `top_k` (top schema candidates retained).

Few-shot retrieval: `tfidf_f`/`emb_f` (sparse vs. dense weights); `fs_k` (retrieved demonstrations inserted).

Voting / control: `agree` (minimum agreement to accept a (*concept_id*, *value*) pair); `margin` (minimum separation between top and runner-up support needed to go to adjudication); `router` (per-transcript budget after routing).

Models: `committee` (LLM models used for extraction and voting); `judge` (adjudication model for conflicts).

L2D-Clinical: Learning to Defer for Adaptive Model Selection in Clinical Text Classification

Rishik Kondadadi, John E. Ortega
University of Minnesota, Northeastern University
Minneapolis, Boston, USA
konda052@umn.edu, j.ortega@northeastern.edu

Abstract

Clinical text classification requires choosing between specialized fine-tuned models (BERT variants) and general-purpose large language models (LLMs), yet neither dominates across all instances. We introduce Learning to Defer for clinical text (L2D-Clinical), a framework that learns when a BERT classifier should defer to an LLM based on uncertainty signals and text characteristics. Unlike prior L2D work that defers to human experts assumed universally superior, our approach enables adaptive deferral—improving accuracy when the LLM complements BERT. We evaluate on two English clinical tasks: (1) ADE detection (ADE Corpus V2), where BioBERT (F1=0.911) outperforms the LLM (F1=0.765), and (2) treatment outcome classification (MIMIC-IV with multi-LLM consensus ground truth), where GPT-5-nano (F1=0.967) outperforms ClinicalBERT (F1=0.887). On ADE, L2D-Clinical achieves F1=0.928 (+1.7 points over BERT) by selectively deferring 7% of instances where the LLM's high recall compensates for BERT's misses. On MIMIC, L2D-Clinical achieves F1=0.980 (+9.3 points over BERT) by deferring only 16.8% of cases to the LLM. The key insight is that L2D-Clinical learns to selectively leverage LLM strengths while minimizing API costs.

Keywords: learning to defer, clinical text classification, model selection, adverse drug events, hybrid AI systems

1. Introduction

Clinical text classification is fundamental to healthcare NLP applications, from adverse drug event (ADE) detection for pharmacovigilance (Henry et al., 2020; Karimi et al., 2015) to treatment outcome extraction for clinical decision support (Xu and Wang, 2022). The choice between specialized fine-tuned models (e.g., BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019)) and general-purpose large language models (LLMs) is often task-dependent, with neither approach universally superior.

Domain-adapted BERT models (Lee et al., 2020; Alsentzer et al., 2019) achieve strong performance when fine-tuned on specific tasks, benefiting from in-domain pretraining. LLMs offer complementary capabilities through broad training and few-shot learning, but their cost, latency, and variable performance across tasks can be limiting factors.

We propose **Learning to Defer for clinical text** (L2D-Clinical), which trains a model to predict when BERT is likely to make errors and should defer to an LLM. Unlike fixed confidence thresholds, L2D-Clinical learns from multiple signals including prediction uncertainty, text characteristics, and domain-specific features.

This setting differs fundamentally from prior L2D work, which defers to human experts assumed *universally superior* when consulted. In our framework, the LLM "expert" may be *worse overall* than BERT (F1=0.765 vs. 0.911 on ADE). Yet L2D-Clinical still improves accuracy by learning to identify the *specific instances* where the weaker-overall model excels. This requires learning fine-grained complementarity rather than simply deferring when uncertain.

To validate this approach, we evaluate on two clinical NLP tasks: (1) **ADE Detection** using the ADE Corpus V2 (Gurulingappa et al., 2012), where BioBERT outperforms the LLM but selective deferral still helps, and (2) **Treatment Outcome Classification** using MIMIC-IV discharge summaries (Johnson et al., 2023),[1] where GPT-5-nano outperforms ClinicalBERT and L2D-Clinical learns optimal deferral patterns.

These pairings reflect the distinct strengths of each model type: domain-adapted BERT models (BioBERT, ClinicalBERT) excel at pattern matching through specialized pretraining, while LLMs leverage broad world knowledge for contextual reasoning. To illustrate the complementarity we aim to exploit, consider a clinical sentence: *"Patient developed rash after starting amoxicillin."* BioBERT, trained on biomedical text, correctly identifies the ADE relationship through learned lexical patterns. However, for *"The medication was well tolerated with no significant issues noted during the hospital stay,"* BioBERT's uncertainty is high because the implicit negation and hedging language fall outside its learned patterns. In such cases, deferring to an LLM that better captures semantic nuance improves accuracy. L2D-Clinical learns to make this

[1] https://physionet.org/content/mimiciv/

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 38–48
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

routing decision automatically, combining BERT's prediction confidence with text features to identify when deferral is warranted.

While the individual components (BERT classifiers, LLM prompting, logistic regression) are well-established, our contribution lies in their novel combination: applying the L2D framework to AI-to-AI deferral in clinical NLP, where the "expert" is not universally superior—a setting not studied in prior work. Our contributions are:

1. **AI-to-AI deferral**: Unlike prior L2D work that defers to human experts, we study deferral between AI systems (BERT $\rightarrow$ LLM), where the "expert" is not universally superior. For example, our LLM achieves only F1=0.765 on ADE detection vs. BERT's 0.911, yet selective deferral still improves overall performance.

2. **Adaptive deferral behavior**: We demonstrate that L2D-Clinical adapts to different model dynamics—improving accuracy when deferral helps (ADE: +1.7 F1 points, MIMIC: +9.3 F1 points) while minimizing LLM usage (7% and 16.8% deferral rates respectively).

3. **Consensus ground truth for MIMIC**: We introduce a multi-LLM consensus labeling methodology (GPT-5.2 + Claude 4.5 agreement) that produces higher-quality ground truth than single-model annotation, with 83.8% agreement rate yielding 2,782 reliable labels. Human expert validation on a 5% subset confirmed 100% label accuracy.

4. **Clinical notes at scale**: Experiments on real MIMIC-IV discharge summaries (279 test samples) alongside the ADE benchmark (500 samples), demonstrating applicability to actual EHR data.

5. **Interpretable patterns**: Analysis reveals task-specific features (e.g., causal language, outcome keywords) that predict when each model excels, supporting clinical validation.

2. Related Work

Our work draws on three research threads: clinical text classification methods, the learning to defer framework, and hybrid NLP systems that combine multiple models for efficient inference.

2.1. Clinical Text Classification

The extraction of ADEs from clinical text has been extensively studied through shared tasks and benchmark datasets. The n2c2 2018 shared task (Henry et al., 2020) established benchmarks for medication and ADE extraction from EHRs, with top systems achieving F1 scores around 0.94 for medication extraction but lower scores for ADE relation extraction. The ADE Corpus V2 (Gurulingappa et al., 2012) provides 23,516 sentences from PubMed case reports annotated for drug-ADE relationships, enabling sentence-level classification research. Karimi et al. (2015) survey text mining techniques for ADE detection, highlighting the challenges of implicit relationships and domain-specific vocabulary.

Treatment outcome extraction from clinical notes is an emerging area with applications in drug repurposing and pharmacovigilance (Xu and Wang, 2022). While structured outcome data exists in resources like clinical trial registries, extracting treatment effectiveness from narrative discharge summaries remains challenging due to implicit language, hedging, and context dependence.

Recent systematic reviews (Ullah et al., 2025) document rapid adoption of LLMs for clinical text classification, though domain-specific BERT models remain competitive on many benchmarks. Zhang et al. (2025) systematically compared BERT and LLM approaches across six clinical and biomedical datasets, finding that "BERT-like models excel in pattern-driven tasks, while LLMs dominate those requiring deep semantics." They propose TaMAS, a task-level selection heuristic based on dataset characteristics. Our work extends this insight: rather than selecting *one* model per task, we learn to select per-instance, achieving gains even when overall model rankings are clear. Agrawal et al. (2022) demonstrate that LLMs can serve as effective few-shot clinical information extractors, though with variable performance across entity types.

2.2. Learning to Defer

The observation that different models excel on different instances motivates the learning to defer framework, where a model learns when to defer decisions to an alternative system. Madras et al. (2018) introduced the framework in the context of fairness, showing that selective deferral can improve both accuracy and demographic parity. Mozannar and Sontag (2020) established theoretical foundations for consistent L2D training, proving that certain surrogate losses yield Bayes-optimal deferral policies. Mozannar and Sontag (2023) provide a comprehensive survey of L2D methods, categorizing approaches by their deferral mechanisms and theoretical guarantees. Raghu et al. (2019) analyze the algorithmic automation problem, examining how to optimally allocate tasks between humans and machines.

In the medical domain, Charusaie et al. (2022) proposed Learning to Defer with Uncertainty (LDU) for clinical diagnosis, achieving F1=0.96 on pleural

effusion detection while deferring 36% of cases to radiologists. Their work assumes the human expert is *always more reliable when deferred to*, a key difference from our setting. Varshney et al. (2022) investigate selective prediction across clinical and general NLP tasks, finding that calibrated uncertainty estimates are crucial for effective abstention.

Our work differs from prior L2D research in three key ways. First, we study **AI-to-AI deferral** where the "expert" (LLM) is not universally superior; on ADE detection, our LLM achieves only F1=0.765 versus BERT's 0.911. Second, our evaluation demonstrates that L2D improves accuracy regardless of which individual model is stronger, validating the approach in both directions. Third, we target **text classification** rather than image-based diagnosis, where the complementary strengths of pattern-matching (BERT) versus semantic reasoning (LLM) create different deferral dynamics than human-AI collaboration.

2.3. Hybrid NLP Systems

Our approach is also related to stacked generalization (Ting and Witten, 1997), which trains a meta-learner to combine base model outputs. However, L2D-Clinical differs in that it routes to one model rather than blending predictions, and its deferral decision is interpretable. Schwartz et al. (2020) propose instance-level model selection based on complexity, showing that matching model capacity to instance difficulty improves efficiency. Swayamdipta et al. (2020) introduce dataset cartography, mapping training dynamics to identify easy, ambiguous, and hard instances—a complementary perspective to our uncertainty-based deferral. Our work operationalizes similar intuitions in a clinical setting, learning per-instance routing between architecturally different models.

Beyond the L2D framework, recent work addresses efficient inference through model cascading and routing. FrugalGPT (Chen et al., 2023) cascades multiple LLMs, using smaller models first and escalating to larger ones when confidence is low. They report up to 98% cost reduction with comparable accuracy on general NLP benchmarks. Our approach differs: (1) we combine *architecturally different* models (BERT vs. LLM) rather than LLM variants, (2) we use learned deferral rather than confidence cascading, and (3) we evaluate on clinical text where domain adaptation matters. Our 83–93% LLM cost reduction with F1 improvements demonstrates L2D-Clinical's effectiveness for specialized domains.

3. Method

Building on these foundations, we develop a lightweight deferral model that routes inputs between BERT and the LLM. The workflow proceeds as follows: (1) BERT classifies the input text and produces softmax probabilities; (2) the deferral model examines these probabilities along with text characteristics to predict whether BERT is likely wrong; (3) if the predicted error probability exceeds a threshold, we query the LLM instead. This routing decision is learned from data, specifically by training on cases where BERT made errors on a validation set. The key insight is that BERT's uncertainty signals, combined with task-specific text features, reliably predict when deferral will help.

3.1. Problem Formulation

Let x be an input text and $y \in \mathcal{Y}$ be the label (binary for ADE detection, 3-class for treatment outcomes). We have access to two classifiers: $f_B(x)$, a domain-adapted BERT classifier (BioBERT for ADE, ClinicalBERT for treatment outcomes), and $f_L(x)$, an LLM with few-shot prompting.

Our goal is to learn a deferral function $d(x) \in [0, 1]$ that outputs the probability of deferring to the LLM. Note that d takes features derived from both the input text x and BERT's output probabilities $f_B(x)$; we write $d(x)$ for brevity. Given threshold τ, the final prediction is:

$$\hat{y} = \begin{cases} f_L(x) & \text{if } d(x) \geq \tau \\ f_B(x) & \text{otherwise} \end{cases} \tag{1}$$

To illustrate, consider two sentences at inference time. For *"Symptoms improved after switching to lisinopril,"* BERT predicts "no ADE" with high confidence; the deferral model outputs $d(x) = 0.12 < \tau$, so we use BERT's prediction. For *"Patient reported feeling dizzy, possibly related to the new medication,"* BERT is uncertain; the deferral model outputs $d(x) = 0.73 \geq \tau$, triggering deferral to the LLM, which correctly identifies the potential ADE.

3.2. Deferral Feature Engineering

The deferral model needs informative features to predict when BERT will fail. We draw on two intuitions: (1) BERT's own uncertainty (reflected in its softmax distribution) signals difficult cases, and (2) certain text characteristics correlate with model-specific strengths. For instance, BERT handles explicit causal language well ("caused by," "due to"), while LLMs better capture implicit relationships and hedged language.

We engineer features that capture when BERT might benefit from LLM assistance:

Uncertainty Features We derive three features from BERT's softmax probability distribution p: (1) *confidence*, defined as $\max_i p_i$, representing the model's certainty in its top prediction; (2) *entropy*, computed as $H(p) = -\sum_i p_i \log p_i$, which captures overall uncertainty across all classes; and (3) *margin*, the difference between the top two class probabilities, indicating decision boundary proximity.

Text Features We include task-specific linguistic signals that correlate with model-specific strengths. Length features capture text complexity via $\log(1 + \text{len})$ for both character and word counts. For ADE detection, we include binary indicators for causal language ("induced," "caused by," "due to," "after," "following"), severity terms ("severe," "fatal"), and common ADE terminology ("toxicity," "reaction," "syndrome"). For treatment outcomes, we include indicators for outcome-related terms ("discontinued," "improved," "intolerance"). These 18 features (5 uncertainty + 13 text) form the input to the deferral model.

3.3. Deferral Model Training

We train the deferral model to predict when BERT is wrong. Let $e(x) = \mathbf{1}[f_B(x) \neq y]$ indicate whether BERT erred on instance x. We minimize binary cross-entropy:

$$d^* = \arg\min_d \sum_x \mathcal{L}_{\text{BCE}}(e(x), d(x)) \tag{2}$$

where $\mathcal{L}_{\text{BCE}}(e, d) = -e \log d - (1 - e) \log(1 - d)$ is binary cross-entropy.

We use logistic regression with balanced class weights to handle the low error rate (BERT is usually correct). The threshold τ is tuned on validation data to maximize the combined system's F1 score. To prevent overfitting, we employ cross-validation: 5-fold for ADE, train/val split for MIMIC-IV.

3.4. Design Choice Justification

Several design decisions warrant explicit justification:

Why Logistic Regression? We prioritize interpretability and simplicity over marginal accuracy gains. Logistic regression provides (1) interpretable coefficients that reveal which features drive deferral, (2) well-calibrated probability estimates suitable for threshold tuning, and (3) minimal overfitting risk given our moderate training set sizes. We evaluated random forest and gradient boosting alternatives; random forest achieved identical F1 on ADE (0.928), and logistic regression matched this performance with fewer hyperparameters and more interpretable outputs. For clinical deployment, understanding *why* the system defers is as important as *when* it defers.

Why These Features? Our feature set combines two complementary signal types. Uncertainty features (confidence, entropy, margin) capture BERT's self-assessed difficulty—high entropy reliably indicates cases where BERT may fail. Text features capture task-specific linguistic patterns that correlate with model-specific strengths. For ADE, explicit causal language ("caused by," "induced") predicts BERT success because such patterns are well-represented in biomedical pretraining. For treatment outcomes, explicit outcome keywords ("discontinued," "improved") similarly predict ClinicalBERT success. These features emerged from error analysis on validation data; we include only features with consistent predictive signal across cross-validation folds.

4. Experimental Setup

Having established the L2D-Clinical framework, we now describe the experimental design used to evaluate its effectiveness across different model dynamics. The ADE Corpus V2 represents a scenario where the domain-adapted BERT model outperforms the LLM, while MIMIC-IV treatment outcome classification represents the opposite, where the LLM outperforms BERT. Evaluating on both scenarios allows us to test whether L2D-Clinical improves accuracy regardless of which base model is stronger, and whether it can effectively leverage a weaker "expert" model.

4.1. Datasets

ADE Corpus V2 We use the ADE Corpus V2 (Gurulingappa et al., 2012), comprising 23,516 English sentences from PubMed medical case reports annotated for ADE presence (binary classification: 29% positive, 71% negative). We download the corpus from the Hugging Face Hub (`ade-benchmark-corpus/ade_corpus_v2`) and split it into train (70%), validation (15%), and test (15%) sets using stratified random sampling with seed 42. For LLM evaluation, we use 500 test instances due to API cost constraints; BERT is evaluated on the full test set (3,527 instances) and the 500-sample subset for fair comparison with the LLM.

MIMIC-IV Treatment Outcomes We extract drug-disease-outcome triplets from 2,000 English-language MIMIC-IV discharge summaries (Johnson et al., 2023), yielding 3,321 candidate triplets.

Outcomes are classified as EFFECTIVE, ADVERSE, or NEUTRAL (3-class). To ensure high-quality ground truth labels, we employ consensus labeling: each triplet is independently classified by GPT-5.2 and Claude 4.5, and only triplets where both models agree are retained. This yields 2,782 consensus-labeled triplets (83.8% agreement rate), providing more reliable ground truth than single-model annotation. To validate the consensus labels, a domain expert reviewed a random 5% subset (139 samples) and confirmed 100% agreement with the consensus labels, supporting the reliability of this methodology. We use note-level train/validation/test splits (80/10/10) with seed 42, yielding 2,225 training, 278 validation, and 279 test samples. GPT-5-nano serves as the deferral target, distinct from the consensus labeling models.

4.2. Models

BERT Models For ADE detection, we fine-tune BioBERT-base-cased-v1.2 (Lee et al., 2020) with the following hyperparameters: learning rate 5×10^{-5} (Hugging Face Trainer default), batch size 16, maximum sequence length 256, warmup ratio 0.1, weight decay 0.01, and early stopping with patience 2 based on validation F1. Training runs for up to 5 epochs. For treatment outcome classification, we fine-tune Bio_ClinicalBERT (Alsentzer et al., 2019) with learning rate 2×10^{-5}, batch size 32, maximum sequence length 512, warmup ratio 0.1, and dropout 0.1 for up to 10 epochs. All experiments use seed 42 for reproducibility.

LLM We use Azure OpenAI GPT-5-nano with few-shot prompting. For ADE detection, the system prompt is: *"You classify sentences for adverse drug events. Reply with only 1 (contains ADE) or 0 (no ADE)."* We provide 5 few-shot examples covering both positive cases (e.g., "Intravenous azithromycin-induced ototoxicity" $\rightarrow$ 1) and negative cases (e.g., "Methotrexate is commonly used for rheumatoid arthritis" $\rightarrow$ 0). For treatment outcomes, the system prompt is: *"You are a clinical NLP assistant. Based on clinical evidence, classify the treatment outcome into one of three categories: EFFECTIVE, ADVERSE, or NEUTRAL. Reply with just the category name."* We provide 3 few-shot examples, one per class. Full prompts are included in our code repository.

Deferral Model We use logistic regression with balanced class weights, L2 regularization (default $C = 1.0$), and maximum 1000 iterations. Features are standardized using z-score normalization. For ADE, we use 5-fold stratified cross-validation to obtain out-of-fold deferral probabilities; for MIMIC-IV, we train on the validation set and evaluate on the held-out test set.

Baselines We compare L2D-Clinical against several baselines: (1) *Fixed confidence threshold*: defer to LLM when BERT's maximum softmax probability falls below θ; we evaluate $\theta \in \{0.6, 0.7, 0.8, 0.9, 0.95\}$ and report $\theta = 0.95$ as it performed best. (2) *Random deferral*: randomly defer at the same rate as L2D-Clinical (7% for ADE, 16.8% for MIMIC) to test whether L2D-Clinical's improvements come from intelligent routing versus simply using the LLM more often. (3) *Oracle*: always choose the correct model when they disagree, providing an upper bound on deferral-based approaches.

4.3. Evaluation Metrics

We report F1 score (harmonic mean of precision and recall) as our primary metric, computed on the positive class for binary ADE detection and macro-averaged across classes for 3-class treatment outcomes. We also report accuracy and the percentage of instances deferred to the LLM (LLM%). For ADE detection, we use 5-fold cross-validation to obtain out-of-fold deferral probabilities, reducing overfitting risk. Given the moderate test set sizes (500 and 279 samples), we acknowledge that small F1 differences ($<$0.01) may not be statistically significant; our reported improvements of +1.7 points (ADE) and +9.3 points (MIMIC) represent meaningful gains. Future work should include bootstrap confidence intervals for more rigorous significance testing.

5. Results

We present results on both tasks, beginning with ADE detection where BioBERT substantially outperforms the LLM (F1: 0.911 vs. 0.765), followed by treatment outcome classification where GPT-5-nano outperforms ClinicalBERT (F1: 0.967 vs. 0.887). In both cases, L2D-Clinical improves over both individual models by learning when to defer.

5.1. Task 1: ADE Detection (BERT $>$ LLM)

Table 1 presents results on ADE Corpus V2, where BioBERT outperforms the LLM.

Despite the LLM's lower overall F1 (0.765), it achieves high recall (95%) compared to BioBERT's 88% recall. This complementarity is key: BioBERT excels at precision (94%) but misses subtle ADEs, while the LLM catches more ADEs but with more false positives (78% precision).

Method	F1	Acc	LLM%
BioBERT only	0.911	0.948	0%
LLM only	0.765	0.834	100%
Random (7%)	0.903	0.938	7.0%
Fixed $\theta = 0.95$	0.921	0.954	2.2%
L2D (Ours)	**0.928**	**0.958**	7.0%
Oracle (upper bound)	0.945	0.968	2.0%

Table 1: ADE Corpus V2: BioBERT outperforms LLM, but L2D-Clinical improves further by selectively using LLM for uncertain cases. Random deferral at the same rate (7%) yields F1=0.903, confirming L2D-Clinical's gains come from intelligent routing.

Error Analysis. BioBERT's false negatives often involve hedged or implicit language. For example, BioBERT missed: *"Patient noted some stomach discomfort which may be related to the medication."* The hedging ("may be") and indirect phrasing caused low confidence. The LLM correctly identified this as a potential ADE. Conversely, the LLM over-predicted on sentences like *"Medication was discontinued and switched to alternative,"* incorrectly inferring an ADE when the switch was for non-adverse reasons. L2D-Clinical learns to defer hedged cases to the LLM while trusting BioBERT on explicit language, improving F1 by 1.7 points.

Residual Errors. Of the 36 errors remaining after L2D-Clinical deferral, 19 (53%) are cases where *both* models fail—typically involving complex multi-drug interactions or rare ADEs absent from BioBERT's pretraining data (e.g., *"Progressive multifocal leukoencephalopathy developed during natalizumab therapy"*). Another 11 (31%) are false deferral cases where the deferral model incorrectly routes to the LLM, which then introduces a new error. The remaining 6 (16%) are cases where the deferral model fails to flag a BERT error. These patterns suggest that further gains require either a stronger base model for rare entities or an ensemble approach rather than binary routing.

5.2. Task 2: Treatment Outcomes (LLM > BERT)

We now examine treatment outcome classification on MIMIC-IV. With consensus ground truth labels (GPT-5.2 + Claude agreement, validated by human expert review), GPT-5-nano outperforms ClinicalBERT, and L2D-Clinical achieves the best performance by learning optimal deferral. Table 2 presents these results.

In this scenario, GPT-5-nano outperforms ClinicalBERT (F1: 0.967 vs. 0.887). The consensus ground truth labels (GPT-5.2 + Claude agreement), validated by human expert review on a 5% subset, provide reliable evaluation. L2D-Clinical achieves F1=0.980 (+9.3 points over BERT) by selectively deferring 16.8% of cases to GPT-5-nano. Notably, L2D-Clinical outperforms the fixed threshold baseline ($\theta = 0.95$, F1=0.971) while using fewer LLM calls (16.8% vs. 26.2%), demonstrating that learned deferral is more effective than confidence-based thresholds.

Method	F1	Acc	LLM%
ClinicalBERT only	0.887	0.946	0%
GPT-5-nano only	0.967	0.968	100%
Fixed $\theta = 0.95$	0.971	—	26.2%
L2D-Clinical (Ours)	**0.980**	**0.986**	16.8%

Table 2: MIMIC-IV Treatment Outcomes with consensus ground truth: GPT-5-nano outperforms ClinicalBERT on this task. L2D-Clinical achieves the best F1 (0.980) by deferring only 16.8% of cases, outperforming the fixed threshold baseline while using fewer LLM calls.

Error Analysis. ClinicalBERT performs well on explicit outcomes: *"Metformin was effective in controlling blood glucose levels"* is correctly classified as EFFECTIVE. However, it struggles with implicit or contextual outcomes like *"Patient tolerated the chemotherapy regimen and was discharged in stable condition"*, where "tolerated" implies effectiveness without explicit statement. GPT-5-nano captures this semantic nuance more reliably. Interestingly, ClinicalBERT excels when explicit keywords appear: sentences containing "discontinued due to" or "adverse reaction" are classified correctly 96% of the time. L2D-Clinical learns to trust ClinicalBERT on these clear-cut cases, reserving the LLM for ambiguous instances where contextual reasoning is needed.

Residual Errors. L2D-Clinical misclassifies 4 of 279 test instances. Two involve ambiguous outcomes where both models disagree with the consensus label—e.g., a discharge note stating *"medication adjusted due to suboptimal response"* was labeled ADVERSE by consensus but could plausibly be NEUTRAL. One is a false deferral where the LLM overrides a correct ClinicalBERT prediction. The final error is a missed deferral on an implicit outcome. The small number of residual errors limits systematic analysis but suggests that remaining failures concentrate at genuine label ambiguity boundaries.

5.3. Summary: Adaptive Deferral Behavior

A key question is whether L2D-Clinical can adaptively learn when deferral helps versus when it would hurt. Table 3 addresses this directly by comparing results across scenarios with different BERT-LLM dynamics.

Task	BERT	L2D	Δ	LLM%
ADE	0.911	**0.928**	+1.7	7.0%
MIMIC	0.887	**0.980**	+9.3	16.8%

Table 3: L2D-Clinical adapts to different model dynamics, improving F1 in both scenarios. On ADE (BERT>LLM), L2D achieves gains with minimal deferral. On MIMIC (LLM>BERT), L2D defers more frequently.

The results demonstrate L2D-Clinical's adaptive behavior. On ADE, even though the LLM is substantially weaker overall (F1=0.765 vs. BERT's 0.911), L2D-Clinical achieves +1.7 F1 points by selectively deferring only 7% of cases where the LLM's high recall compensates for BERT's misses. On MIMIC, where GPT-5-nano outperforms ClinicalBERT (F1: 0.967 vs. 0.887), L2D-Clinical achieves +9.3 points improvement by deferring 16.8% of cases—learning a higher deferral rate that reflects the LLM's superior performance on this task. This adaptive behavior—adjusting deferral rates based on relative model strengths—demonstrates that L2D-Clinical learns meaningful task-specific patterns rather than applying a one-size-fits-all strategy.

5.4. Feature Analysis

To understand what drives effective deferral decisions, we examine the learned feature weights. Table 4 shows standardized logistic regression coefficients, where positive values indicate features that increase deferral probability. Across both tasks, BERT's prediction entropy is the strongest signal: high entropy reliably indicates cases where deferral helps.

Feature	ADE	MIMIC
Entropy	+0.82	+0.91
Margin (top-2 diff)	−0.71	−0.68
Confidence	−0.54	−0.47
Log(text length)	+0.23	+0.31
Has "discontinued"	—	−0.44
Has causal language	−0.38	—

Table 4: Deferral model coefficients (positive = defer to LLM). Uncertainty features dominate; task-specific keywords add interpretable signal.

Task-specific patterns emerge: explicit causal language ("caused by," "after taking") predicts correct BERT classification on ADE, while "discontinued" keywords predict correct BERT on MIMIC. Ambiguous outcome language ("tolerated," "stable on") triggers deferral. These interpretable patterns enable clinicians to understand and validate deferral decisions.

6. Discussion

The experimental results raise a natural question: why does L2D-Clinical succeed in improving accuracy even when deferring to a weaker model? We now analyze the underlying mechanisms and discuss practical implications for clinical deployment.

6.1. Why L2D-Clinical Works

The central finding, that L2D-Clinical adapts its deferral behavior to model dynamics, warrants deeper examination. On ADE, deferring to a substantially weaker model (LLM F1=0.765) still helps BERT (F1=0.911). The key insight is that aggregate F1 scores mask instance-level variation: a model that performs worse on average can still outperform on specific subsets of data where its strengths align with the input characteristics. On MIMIC, where the LLM outperforms BERT (F1: 0.967 vs. 0.887), L2D-Clinical learns a higher deferral rate (16.8%) to leverage the LLM's strengths, achieving the best overall F1 (0.980).

Table 5 contrasts our setting with prior L2D work.

	Prior L2D	L2D-Clinical
Expert type	Human	AI (LLM)
Expert reliability	Always better	Instance-dep.
Deferral signal	Uncertainty	Uncert. + feat.

Table 5: L2D-Clinical vs. prior Learning to Defer approaches.

L2D-Clinical succeeds because of three key factors:

Factor 1: Complementary Errors. On ADE, BioBERT achieves 94% precision but only 88% recall; the LLM shows the opposite pattern with 78% precision and 95% recall. These complementary error profiles mean that even a weaker-overall model can rescue specific failure cases. BERT misses hedged language; the LLM catches it. The LLM over-predicts on ambiguous cases; BERT correctly abstains.

Factor 2: Learnable Error Prediction. The deferral model achieves 73% precision at predicting

BERT errors on ADE and 68% on MIMIC, meaning that when it predicts BERT will fail, it is correct 73% and 68% of the time respectively. Given that BERT's base error rates are only 12% (ADE) and 18% (MIMIC), this represents substantial predictive power. Even imperfect predictions improve overall accuracy when the complementary model is consulted on flagged cases.

Factor 3: Selective Deferral. By deferring only when confident BERT will fail *and* the LLM is likely correct, L2D-Clinical avoids introducing new errors. On ADE, of the 7% of cases deferred to the LLM, the LLM was correct 89% of the time, much higher than its 83.4% overall accuracy. This selective routing is the key mechanism.

6.2. Cost Analysis

L2D-Clinical provides substantial computational savings by using the LLM for only 7% (ADE) and 16.8% (MIMIC) of instances, reducing LLM API costs by 81–93% compared to LLM-only inference while maintaining or improving accuracy. A detailed cost and latency analysis is provided in Appendix 8.2.

6.3. Practical Implications

L2D-Clinical offers several advantages for clinical deployment:

Model-Agnostic Design. L2D-Clinical works regardless of which model is individually better, making it robust to task-specific model dynamics. Practitioners do not need to know *a priori* whether BERT or LLM will be stronger; L2D-Clinical adapts automatically.

Interpretable Deferral. Unlike black-box ensemble methods, L2D-Clinical's deferral decisions are interpretable. Clinicians can understand *why* a case was deferred: high entropy, presence of hedging language, or absence of explicit keywords. This transparency supports clinical validation and trust.

Graceful Degradation. If the LLM API becomes unavailable or too slow, L2D-Clinical can fall back to BERT-only predictions with known accuracy bounds. This is critical for clinical systems requiring high availability.

Privacy Considerations. The deferral model can be configured to limit LLM usage for sensitive cases, reducing data exposure to external APIs.

7. Conclusion

We have presented Learning to Defer for clinical text (L2D-Clinical), demonstrating that learned deferral adapts to different model dynamics. On ADE detection, where BioBERT outperforms the LLM but complementary error patterns exist, L2D-Clinical achieves F1=0.928 (+1.7 points over BERT alone) by selectively deferring 7% of instances. On MIMIC-IV treatment outcomes with consensus ground truth, where GPT-5-nano outperforms ClinicalBERT (F1: 0.967 vs. 0.887), L2D-Clinical achieves F1=0.980 (+9.3 points over BERT) by deferring 16.8% of cases. The deferral model learns interpretable, task-specific patterns about when deferral helps, providing transparency that supports clinical validation.

The broader implication of this work is that L2D-Clinical provides a principled framework for combining specialized and general-purpose models. Rather than choosing one architecture or blindly deferring when uncertain, L2D-Clinical learns when deferral is likely to help and adjusts deferral rates accordingly. This adaptive behavior is particularly valuable as the landscape of available models evolves: L2D-Clinical can incorporate new models and learn appropriate deferral patterns without manual tuning. Our consensus labeling methodology for MIMIC-IV, combining multi-LLM agreement with human expert validation, demonstrates a practical approach to generating high-quality ground truth for clinical NLP tasks where expert annotation is expensive.

However, significant challenges remain for deploying L2D-Clinical in production clinical settings. The deferral model requires validation data where both systems have been evaluated, creating a bootstrapping problem for new tasks. The interpretability of deferral decisions, while better than black-box ensembles, still requires clinical validation to ensure that the learned patterns align with meaningful medical distinctions rather than spurious correlations. Finally, the rapid pace of LLM development means that deferral patterns learned today may become suboptimal as models improve, necessitating ongoing recalibration.

Data Availability

The ADE Corpus V2 is publicly available from the Hugging Face Hub (`ade-benchmark-corpus/ade_corpus_v2`). MIMIC-IV access requires credentialing through PhysioNet (https://physionet.org/content/mimiciv/).

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 (EMNLP)*, pages 1998–2022.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew BA McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. 2022. Incorporating uncertainty in learning to defer algorithms for safe computer-aided diagnosis. *Scientific Reports*, 12(1):1–12.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Frugalgpt: How to use large language models while reducing cost and improving performance. In *arXiv preprint arXiv:2305.05176*.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Sarvnaz Karimi, Alejandro Metke-Jimenez, An Nguyen, et al. 2015. Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys*, 47(4):1–39.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, 8(12):2100707.

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.

David Madras, Toniann Pitassi, and Richard Zemel. 2018. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31.

Hussein Mozannar and David Sontag. 2020. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR.

Hussein Mozannar and David Sontag. 2023. Teaching models to abstain when uncertain: A survey of learning to defer methods. *Journal of Machine Learning Research*, 24:1–55.

Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. 2019. The algorithmic automation problem: Prediction, triage, and human effort. In *arXiv preprint arXiv:1903.12220*.

Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A Smith. 2020. The right tool for the job: Matching model and instance complexities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6640–6651.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293.

Kai Ming Ting and Ian H Witten. 1997. Stacking bagged and dagged models. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, pages 367–375.

Irfan Ullah et al. 2025. Large language models for healthcare text classification: A systematic review. *arXiv preprint arXiv:2503.01159*.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022. Investigating selective prediction approaches across several tasks in iid, ood, and adversarial settings. In *Findings of ACL*.

Rui Xu and Qingqing Wang. 2022. A survey on adverse drug event detection from texts. *Expert Systems with Applications*, 205:117627.

Yusheng Zhang et al. 2025. Do bert-like bidirectional models still perform better on text classification in the era of llms? In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

8. Appendix

Sections in this appendix provide more insight into the work completed. We discuss the limitations for our infrastructure along with its latency and how we plan on presenting more ideas in the future.

8.1. Limitations

Several limitations of this work should be acknowledged. First, our test set sizes are moderate (500 and 279 samples) due to LLM API cost constraints; larger-scale evaluation would strengthen generalizability claims. Second, we evaluate only a single LLM family (GPT-5); deferral patterns may differ with open-source alternatives such as Llama or Mistral, which have different capability profiles. Testing whether the learned deferral mechanism generalizes across LLM families is an important direction for future work. Third, L2D-Clinical requires labeled validation data to train the deferral model, and cold-start scenarios where such data is unavailable need further investigation. Fourth, for the MIMIC task, the consensus ground truth labels were generated by LLMs (GPT-5.2 and Claude 4.5), while the deferral target is also from the GPT family (GPT-5-nano). Although the labeling and inference models are distinct, this shared lineage could introduce bias favoring LLM-like reasoning patterns. The 5% human expert validation mitigates but does not fully eliminate this concern; larger-scale human validation and evaluation on existing gold-labeled clinical datasets (e.g., n2c2 shared task corpora) would further strengthen confidence. Fifth, we defer exclusively to an LLM, but deferring to a smaller model fine-tuned with emphasis on high recall could potentially achieve similar gains at lower cost—investigating non-LLM deferral targets is a promising alternative. Sixth, our evaluation is English-only; clinical NLP in other languages may exhibit different BERT-LLM dynamics. Finally, when using cloud-based LLMs, deferral sends patient data to external APIs; deployments involving protected health information (PHI) require either on-premise LLM deployment or Business Associate Agreement (BAA) compliance with the API provider.

8.2. Cost and Latency Analysis

Beyond the accuracy improvements, L2D-Clinical provides substantial computational savings, a critical consideration for deploying clinical NLP at scale. We analyze cost and latency assuming LLM inference costs 50$\times$ BERT per call (based on typical API pricing) and 850ms latency versus 12ms for BERT. On the ADE task, L2D-Clinical uses only 7% LLM calls, resulting in 4.4$\times$ relative cost compared to 50$\times$ for LLM-only inference, with an average latency of 71ms per sample. On the MIMIC task, L2D-Clinical uses 16.8% LLM calls, yielding 9.4$\times$ relative cost with 153ms average latency.

These estimates reflect only LLM API inference costs. The full computational cost also includes BERT fine-tuning (approximately 20 GPU-minutes on a single A100 for BioBERT on ADE) and L2D classifier training (under 1 CPU-second for logistic regression on 18 features). At inference time, the L2D classifier adds $<$1ms overhead per sample. Using the Green Algorithms calculator (Lannelongue et al., 2021), we estimate total training carbon emissions at approximately 0.03 kg CO_2e for BERT fine-tuning and negligible emissions for logistic regression, compared to the ongoing emissions from full LLM inference at scale.

For high-throughput clinical pipelines processing thousands of notes daily, these savings are significant. A hospital processing 10,000 clinical notes per day would reduce LLM API costs by 81–93% compared to LLM-only inference, while maintaining or improving accuracy.

8.3. Future Work

While the present work focused on classification tasks, the L2D-Clinical framework opens several promising research directions:

Other Clinical NLP Tasks. Named entity recognition, relation extraction, and temporal reasoning may exhibit similar complementary error patterns between specialized and general models. For NER, BERT-based taggers excel at common entities but struggle with rare or context-dependent mentions where LLMs might help.

Multi-Model Deferral. Extending beyond two models to a cascade of increasingly capable (and expensive) models could further optimize the accuracy-cost tradeoff. The deferral model would learn when to escalate through the cascade.

Online Adaptation. As LLM capabilities evolve rapidly, deferral thresholds may need periodic recalibration. Investigating online learning approaches that adapt to distribution shift without full retraining is important for long-term deployment.

Cross-Lingual Transfer. Clinical NLP in languages other than English may show different

BERT-LLM dynamics. Evaluating whether L2D-Clinical patterns transfer across languages would broaden applicability.

TRUMEDIQA: A Modular Trustworthy RAG Pipeline for Multilingual Medical Question Answering

Meryem EL Fatimi[1], **Ayoub Nainia**[2], **Jihad Zahir**[3]
[1] TICLab, International University of Rabat, Rabat, Morocco
[2] Institut de Systématique, Évolution, Biodiversité (ISYEB), Sorbonne Université, Paris, France
[3] UMMISCO, IRD, Computer Science Department, LISI, Cadi Ayyad University,
Faculty of Sciences, Marrakech, Morocco
meryem.elfatimi@uir.ac.ma, ayoub.nainia@sorbonne-universite.fr, j.zahir@uca.ac.ma

Abstract

Medical question answering systems must balance usefulness with safety, particularly in low-resource linguistic settings where robustness is limited and hallucinations can cause harm. We present TRUMEDIQA, a reproducible multilingual medical QA pipeline for Moroccan Darija, Arabic, French, and English, deployed on WhatsApp with text and voice interactions. TRUMEDIQA uses layered decision-making: (i) language identification, (ii) a pre-retrieval intent router that maps queries to one of 38 clinical FAQ categories to constrain retrieval, and (iii) post-retrieval LLM-based re-ranking that selects the best candidate answer or returns a null decision to trigger a safe fallback (abstention). Answers are retrieved from a curated FAQ knowledge base validated by medical professionals. We evaluate TRUMEDIQA with 21 participants submitting 290 questions across four languages. An expert annotator labels each interaction as relevant, acceptable, or irrelevant, and we also measure correct abstentions when no suitable answer exists in the knowledge base. An ablation study shows that routing and re-ranking improve the weighted relevance score from 0.25 to 0.94 and precision from 0.53 to 0.98 versus a naïve retrieval baseline, while increasing correct abstention on unanswerable queries from 4.38% to 69.77%.

Keywords: clinical question answering, multilingual NLP, retrieval-augmented generation

1. Introduction

Large Language Models (LLMs) have recently improved the fluency and apparent competence of medical conversational systems, creating opportunities to expand access to health information at scale. However, medical QA remains high-risk: models can produce plausible but incorrect statements, and this risk is amplified in low-resource linguistic settings where robustness is limited. In the Arabic context, recent benchmarks show that even Arabic-centric LLMs perform poorly on dialectal input compared to Modern Standard Arabic (MSA) (Mousi et al., 2025). Work targeting Moroccan Arabic (Darija) similarly reports that substantial adaptation is required to handle dialectal variation reliably (Qarah and Alsanoosy, 2025). These gaps can translate into unequal access to safe medical information and motivate clinical NLP systems that are both multilingual and cautious under uncertainty.

In this work, we study patient-facing medical question answering for adolescent reproductive health in a multilingual setting (Moroccan Darija, Arabic, French, and English). We introduce TRUMEDIQA, a reproducible WhatsApp-based QA system that combines retrieval with layered decision-making to reduce unsafe mismatches and to abstain when the knowledge base lacks a suitable answer. TRUMEDIQA retrieves answers from a curated FAQ knowledge base validated by medical professionals, and applies (i) language identification, (ii) a pre-retrieval intent router over 38 clinical FAQ categories to constrain retrieval, and (iii) post-retrieval LLM-based re-ranking that selects the best candidate answer or returns a null decision to trigger a safe fallback. This design aims to improve reliability without overclaiming coverage, which is critical in medical settings. TRUMEDIQA is intended for informational support rather than diagnosis or treatment advice. While individual components such as intent routing, constrained retrieval, and abstention are established patterns in RAG systems, their integration in a real-world multilingual deployment targeting a low-resource dialect (like Moroccan Darija) for a socio-culturally sensitive health domain represents a novel and underexplored setting that existing systems do not address.

This paper makes the following contributions:

- We present TRUMEDIQA, a layered multilingual medical QA system supporting Moroccan Darija, Arabic, French, and English, with both text and voice interactions on WhatsApp.
- We propose a modular decision pipeline combining intent-based query routing and post-retrieval LLM re-ranking with explicit abstention to reduce unsafe mismatches.
- We report a real-world evaluation (21 participants, 290 questions) with expert annotation, and an ablation study quantifying the impact of

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 49–56
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

routing and re-ranking on relevance, precision, and correct abstentions.

- We provide system details and a reproducible evaluation protocol (taxonomy, prompts, configuration), and describe our privacy/anonymization handling to facilitate replication in other low-resource health settings.

- We demonstrate that a safety-oriented layered RAG architecture can be effectively deployed in a low-resource dialectal setting, addressing a gap in multilingual clinical NLP where Moroccan Darija and similar dialects remain severely underserved.

The remainder of this paper is structured as follows: Section 2 reviews related work on medical QA and clinical NLP. Section 3 describes the trustworthiness dimensions that guide our design. Section 4 presents the TRUMEDIQA pipeline and knowledge base construction. Section **??** reports the experimental setup, and Sections 6 and 7 define the evaluation framework and metrics. Results are reported in Section 8, followed by discussion in Section 9 and conclusion in Section 10. Finally, we provide an ethics statement and limitations in Sections 11 and 12

2. Background and Related Work

2.1. Patient-facing medical QA and multilingual gaps

Medical question answering (MQA) aims to return reliable health information in response to natural-language queries. Prior to LLMs, many MQA systems relied on modular pipelines that combined question classification, retrieval over curated resources, and answer selection, often leveraging controlled terminologies such as UMLS and SNOMED CT for normalization and interpretability (Mutabazi et al., 2021; Zhu et al., 2018). While modern LLMs have improved fluency and coverage, reliability and safety remain central concerns in patient-facing settings.

A major limitation is linguistic imbalance: most benchmarks and resources are English-centric, which yields systematic performance gaps for other languages (Hamad et al., 2024; Baethge, 2008; Valkimadi et al., 2009). Multilingual medical QA benchmarks such as MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and MultiMedQA (Singhal et al., 2025) show strong results on structured exam-style questions, yet performance degrades on open-ended queries and in linguistically diverse settings. Recent evaluations (e.g., MedExpQA) further report persistent disparities between English and non-English answers, even under translation (Alonso et al., 2024). These findings motivate multilingual systems that explicitly account for low-resource languages and dialectal variation.

2.2. Grounded LLM-based medical QA and retrieval pipelines

To mitigate hallucinations and improve factuality, a common strategy is to ground generation in external knowledge via retrieval-augmented generation (RAG) (Lewis et al., 2020). In medical QA, RAG typically retrieves from trusted sources (e.g., guidelines, curated FAQs, biomedical literature) and then generates or selects answers conditioned on retrieved evidence. Recent work explores retrieval and ranking improvements, including domain-specific retrievers and medically oriented RAG systems (Zhao et al., 2025; Gao et al., 2025). TRUMEDIQA follows this line but emphasizes a layered decision pipeline that combines intent-based routing and post-retrieval re-ranking to reduce mismatches in multilingual user queries.

2.3. Safety under uncertainty: calibration, answerability, and abstention

Even when grounded, LLMs can be overconfident and may produce misleading outputs (Tian et al., 2023; Nazi and Peng, 2024). Calibration and uncertainty estimation methods (e.g., temperature scaling, ensembles, and self-assessment prompts) aim to better align model confidence with correctness (Desai and Durrett, 2020; Balabanov and Linander, 2025; Xiong et al., 2024). In high-stakes medical contexts, this motivates *selective answering*: abstaining when evidence is insufficient or when a query is outside the supported scope. Our design operationalizes this principle through explicit abstention triggered by routing and re-ranking decisions, enabling the system to defer when the FAQ knowledge base does not contain an appropriate answer.

3. Trustworthiness by design

We define trustworthiness for multilingual medical QA as a set of system properties that can be operationalized and audited in the pipeline. In TRUMEDIQA, trustworthiness relies on four dimensions: (1) accuracy and safety, (2) explainability, (3) compliance with local socio-cultural and legal constraints, and (4) privacy. Table 1 summarizes how each dimension is instantiated in the system and how it is assessed in our evaluation.

The remainder of this paper focuses on the system components that instantiate these dimensions and on a real-world evaluation quantifying the trade-off between answer quality and responsible abstention.

4. Medical Knowledge Base Construction

TRUMEDIQA retrieves answers from a curated FAQ knowledge base covering adolescent reproductive health. The knowledge base is designed for patient-facing informational support and is intentionally scoped to avoid diagnosis, treatment recommendations, or urgent-care guidance; out-of-scope or emergency-like queries trigger abstention and a redirect message (Section 3).

FAQ authoring and clinical validation. We first compile a set of recurring user questions for the target domain, initially authored in Moroccan Darija. For each question, a medical professional authored and/or validated the corresponding answer to ensure local appropriateness and consistency with the intended scope. Answers may include optional media pointers (e.g., links or simple visual aids) when useful for comprehension.

Multilingual collections. To support multilingual access, we build separate FAQ collections for Moroccan Darija, Arabic, French, and English. Each collection contains language-specific question formulations paired with clinician-validated answers written for that language. This design avoids relying on cross-lingual retrieval as a default and enables language-aware routing and evaluation within each collection.

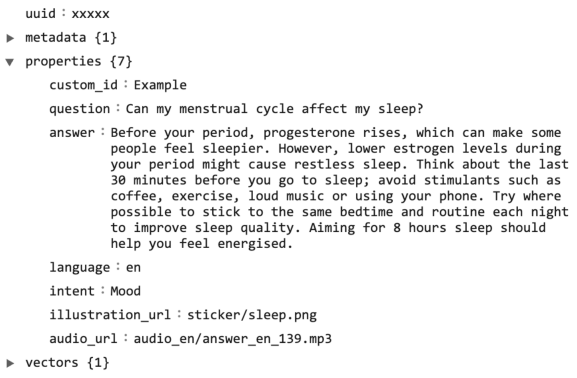


Figure 1: TRUMEDIQA Knowledge Base Entry.

Intent taxonomy. Each FAQ is manually assigned an intent label from a fixed taxonomy of 38 clinical FAQ categories. These labels provide an interpretable organization of the knowledge base and enable intent-based query routing that constrains retrieval to category-consistent candidates.

Vector indexing and stored schema. We store each language collection in Weaviate as a dedicated vector index. Each entry contains the question text, a clinician-validated answer, a language identifier, an intent category from our 38-category taxonomy, and optional media fields (e.g., `audio_url`, `sticker_url`; Figure 1). Within a given language, we embed question texts into dense vectors for similarity search and retrieve top-k candidates at inference time, which are then

Dimension	Operationalization in TRUMEDIQA	Evidence / assessment
Accuracy & safety	Clinician-validated FAQ knowledge base; intent-based routing to restrict retrieval space; post-retrieval re-ranking; explicit abstention when no suitable answer is found; safety-oriented fallback messages.	Expert annotation (relevant/acceptable/irrelevant); Precision and weighted Relevance Score; abstention analysis (Fallback Rate and correct abstentions).
Traceability (grounding)	Responses are linked to a retrieved FAQ entry; the system can surface the matched FAQ identifier and/or a supporting snippet; explicit abstention for uncovered queries.	Grounding coverage (% responses linked to an FAQ ID/snippet) and qualitative inspection; abstention correctness captured by correct-abstention metrics.
Local compliance	Domain scoping to adolescent reproductive health; standardized disclaimers and response templates; constrained answer set to avoid unavailable/inappropriate recommendations.	Qualitative analysis of compliance-related cases; failure modes reported in limitations.
Privacy	Anonymized logging; data minimization; restricted access to interaction logs; raw voice messages not retained beyond processing in our deployment.	Description of consent/anonymization and data handling; deployment constraints and limitations.

Table 1: Trustworthiness dimensions and their operationalization in TRUMEDIQA.

passed to a post-retrieval re-ranking stage.

Re-ranking module. Retrieved top-k candidates are passed to an LLM-based re-ranker for final selection. The re-ranker uses GPT-4o and receives a structured prompt containing the user question, the detected intent category, and the list of candidate question–answer pairs indexed from 0 to $k-1$. The system prompt instructs the model to act as an answer ranker rather than a generator: it must return the integer index of the most semantically relevant candidate, or `null` if no candidate adequately addresses the question. The model is constrained to respond with a single token only — an integer or the word `null`. A `null` response triggers the safe fallback mechanism described in Section 3. Temperature is set to 0 to ensure deterministic and reproducible ranking decisions.

Voice responses. To reduce latency for voice interactions, we pre-generate an audio rendition for each validated answer using a text-to-speech engine and store it as an `audio_url` field. When an answer is selected, TRUMEDIQA can return the text response and, when applicable, the corresponding audio response.

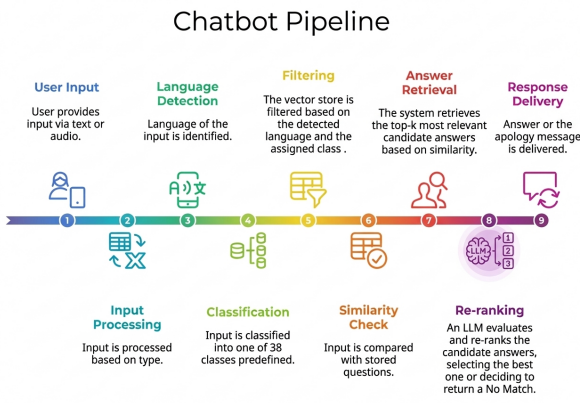


Figure 2: Overview of the TRUMEDIQA chatbot pipeline, from user input to response delivery. The pipeline combines language detection, intent-based classification, filtered retrieval, LLM re-ranking, and safe fallback decision process.

5. Experimental Setup

We evaluated TRUMEDIQA with 21 volunteers, who submitted 290 questions during a four-hour testing window via WhatsApp. Participants could ask questions in Arabic, Moroccan Darija, French, or English, and could freely alternate between languages. The language distribution reflects the target population: 220 questions were submitted in Arabic or Moroccan Darija — which participants used interchangeably and represent the primary language of our deployment context — 37 in English, 30 in French, and 3 involved code-switching across languages. This predominance of Arabic/Darija is consistent with our motivation: Moroccan users naturally favor Darija, a low-resource dialect, making it both the most representative and the most challenging language in our evaluation. Inputs were provided either as text messages or as voice messages (up to 60 seconds); voice messages were transcribed using Google Cloud Speech-to-Text before being passed to the pipeline. All interactions were anonymized for analysis and collected with participant consent; the system is intended for informational support rather than diagnosis or treatment advice.

We conducted an ablation study to quantify the contribution of two key components: intent routing (38-category taxonomy) and post-retrieval re-ranking. We compared three configurations (Table 3) on the same set of user questions. C1 is a naive retrieval baseline without routing or re-ranking. C2 adds intent routing but removes re-ranking, isolating the effect of routing. C3 corresponds to the complete system.

6. Evaluation Framework

All interactions were logged and anonymized to protect participant privacy. After the four-hour testing window, an expert evaluator manually annotated each question-answer (QA) pair. The evaluation focuses on whether the system took the *correct action* given the knowledge base coverage: answering when an entry exists and abstaining when it does not. Each interaction was assigned to one of three outcome classes:

- **Relevant:** The system took the correct action. This includes:
  - **Correct answer (answerable):** a direct,

Config.	System variant	What it tests
C1	Direct vector retrieval (no routing, no re-ranking).	Baseline retrieval quality without filtering.
C2	Intent routing (38 categories) + retrieval; no re-ranking.	Contribution of routing to candidate quality.
C3	Full system: routing + retrieval + LLM re-ranking (+ abstention).	End-to-end performance of TRUMEDIQA.

Table 2: Ablation configurations for evaluating routing and re-ranking.

correct, and contextually appropriate answer when a suitable FAQ entry exists.

  - **Correct abstention (unanswerable):** abstention via a fallback message when no suitable answer exists in the knowledge base.

- **Acceptable:** The response was topically relevant and potentially useful but partially correct, incomplete, or insufficiently specific.
- **Irrelevant:** The response exhibited a clear mismatch between the question and the returned answer, or the system abstained despite the existence of a suitable knowledge base entry (i.e., incorrect abstention).

7. Evaluation Metrics

To assess TRUMEDIQA, we report four complementary metrics: (i) a weighted Relevance Score summarizing expert quality labels, (ii) precision over answered queries to quantify the rate of misleading answers, (iii) fallback rate to characterize abstention frequency, and (iv) abstention correctness to distinguish appropriate abstention from unnecessary fallback. Together, these metrics capture both answer quality and the system's ability to manage uncertainty.

Relevance Score. We compute a weighted relevance score (RS) from expert labels by assigning weights 2/1/0 to *Relevant*/*Acceptable*/*Irrelevant* and normalizing to $[0, 1]$:

$$RS = \frac{1 \cdot N_{\text{acceptable}} + 2 \cdot N_{\text{relevant}}}{2 \cdot N_{\text{total}}}. \quad (1)$$

Precision (answered queries). We define an *answered* query as one where the system returns an explicit answer rather than a fallback message. A false positive (FP) is an answered query labeled *Irrelevant* by the expert. A true positive (TP) is an answered query labeled *Relevant* or *Acceptable*. Precision is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (2)$$

Fallback Rate. Fallback Rate (FR) measures how often the system abstains:

$$\text{FR} = \frac{N_{\text{fallback}}}{N_{\text{total}}}. \quad (3)$$

Abstention correctness. To evaluate whether abstentions are appropriate given knowledge base coverage, the expert additionally labels each query as *Answerable* or *Unanswerable* with respect to the FAQ knowledge base. Correct abstentions (true negatives, TN) are fallback cases labeled *Unanswerable*:

$$TN = \sum_{i=1}^{N} \mathbf{1}(\hat{a}_i = \text{Fallback}) \cdot \mathbf{1}(a_i = \text{Unanswerable}), \quad (4)$$

where $\hat{a}_i$ is the system action (answer vs fallback) and a_i is the expert answerability label. Conversely, incorrect abstentions (false negatives, FN) are fallback cases labeled *Answerable*:

$$FN = \sum_{i=1}^{N} \mathbf{1}(\hat{a}_i = \text{Fallback}) \cdot \mathbf{1}(a_i = \text{Answerable}). \quad (5)$$

We report TN as a rate by normalizing by $N_{\text{unanswerable}}$ and FN as a rate by normalizing $N_{\text{answerable}}$ (FNR)..

Configuration	RS	Prec.	TN (%)	FR (%)
C1: no routing, no re-ranking	0.25	0.53	4.38	55
C2: routing only (no re-ranking)	0.68	0.74	26.47	12
C3: full system	**0.94**	**0.98**	**69.77**	15

Table 3: Ablation results. TN and FR are reported as percentages. TN(%) denotes correct abstentions among unanswerable queries, computed as $TN/N_{\text{unanswerable}}$. FR(%) denotes the overall fallback rate, computed as $(TN + FP)/N$.

8. Results

Table 3 reports the ablation results for routing and re-ranking. The naive baseline (C1) yields low overall quality (RS=0.25) and moderate precision (0.53), with a high fallback rate (55%), indicating frequent abstention and limited utility in this setting. Adding intent routing (C2) substantially improves RS (0.68) and precision (0.74) while reducing fallback to 12%, suggesting that constraining retrieval to category-consistent candidates reduces mismatches. The full system (C3) achieves the best performance (RS=0.94, precision=0.98) and markedly increases correct abstention (TN; Section 7), indicating improved alignment with knowledge base coverage limits. Figure 3 shows the distribution of expert labels across configurations, where the number of irrelevant responses drops sharply from C1 to C3.

8.1. Error Analysis

To better understand the residual failures of the full system (C3), we manually analyzed all cases labeled Irrelevant or Acceptable by the expert annotator. We identify three failure types.

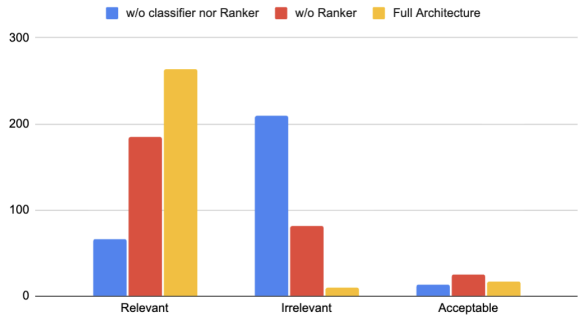


Figure 3: Expert evaluation of all configurations (C1,C2 and C3)

False abstention (14 cases, 52%). The most frequent failure occurs when the ranker returns a null decision despite the existence of a suitable FAQ entry in the knowledge base. This typically arises when the retrieved candidates are semantically close but not close enough for the ranker to commit to a match, leading to an overly conservative abstention. This pattern was most prevalent in Darija queries, where the embedding model produces less reliable similarity scores due to the low-resource nature of the dialect.

Wrong answer selected (8 cases, 30%). In these cases the ranker returns an answer but selects the wrong FAQ entry even though the correct one exists in the knowledge base. This reflects retrieval ambiguity within a correctly identified intent category, where multiple candidates share surface similarity with the query.

Incorrect answer when abstention was appropriate (5 cases, 19%). In these cases the ranker returned an answer for a query that has no suitable match in the knowledge base, when a NO MATCH decision would have been correct. These cases involve questions that fall outside the scope of the knowledge base, where the ranker should have abstained but instead returned a partially related answer.

Overall, false abstentions account for the majority of C3 failures, suggesting that the ranker's conservatism — while beneficial for safety — occasionally withholds valid answers.

9. Discussion

The ablation results (Table 3) show a consistent improvement as routing and re-ranking are added (Figure 3). The weighted relevance score increases from 0.25 in the naive baseline (C1) to 0.94 in the full system (C3), and precision improves from 0.53 to 0.98. Together, these trends indicate that constraining retrieval to category-consistent candidates and re-ranking retrieved entries substantially reduces mismatches and misleading answers.

Beyond answer quality, the results also highlight the role of explicit abstention for safe behavior under limited knowledge base coverage. While the full system exhibits a modest fallback rate (15%), it achieves a substantially higher correct-abstention rate (TN; Section 7) than the baseline. This suggests that TRUMEDIQA more often withholds answers in cases where the knowledge base does not contain a suitable entry, rather than returning low-quality matches.

Overall, the comparison across configurations underscores the complementary value of intent routing and post-retrieval re-ranking in multilingual, low-resource medical QA. Routing improves candidate selection by narrowing the retrieval space, while re-ranking further improves contextual alignment and supports more reliable abstention decisions. These components jointly contribute to a better balance between usefulness and caution, which is critical in patient-facing medical settings.

10. Conclusion

We presented TRUMEDIQA, a reproducible multilingual medical QA pipeline for Moroccan Darija, Arabic, French, and English, deployed on WhatsApp with text and voice interactions. In a real-world evaluation (21 participants, 290 questions), an ablation study showed that combining intent routing with post-retrieval re-ranking substantially improves response quality compared to a naive retrieval baseline (RS: 0.25 to 0.94; precision: 0.53 to 0.98), while supporting responsible abstention when no suitable knowledge base entry exists.

Future work will focus on expanding coverage while preserving safety guarantees. One direction is to extend the curated FAQ knowledge base with additional clinician-reviewed content derived from medical documents, using human-in-the-loop validation for extracted passages. Another direction is to improve uncertainty handling by modeling answerability more explicitly and reducing unnecessary abstentions, as well as enabling controlled knowledge base updates with versioning and periodic clinical review.

11. Ethics Statement

TRUMEDIQA is a patient-facing informational QA system designed to provide general health education within a constrained scope (adolescent reproductive health). It is **not** intended for diagnosis, treatment recommendation, or emergency triage. To reduce the risk of harm, the system is designed to abstain when knowledge base coverage is insufficient and to present standardized fallback mes-

sages that encourage consulting qualified health professionals when appropriate.

The evaluation involved 21 volunteers interacting with the system via WhatsApp. All interactions were logged for research purposes and anonymized prior to analysis. We followed data minimization principles by restricting access to logs and avoiding the persistent storage of raw voice messages beyond processing. Because reproductive health topics can be sensitive, we emphasize that any deployment of similar systems should include clear user-facing disclosures, consent procedures, and appropriate safeguards for vulnerable users.

Finally, our work highlights equity considerations in clinical NLP: low-resource languages and dialects (e.g., Moroccan Darija) are under-served by current medical NLP systems, which can exacerbate unequal access to safe health information. Our goal is to contribute methods and evaluation practices that improve reliability for such settings.

12. Limitations

Our study has limitations. First, TRUMEDIQA relies on a curated FAQ knowledge base; therefore, system coverage is bounded by the scope and completeness of this resource. While abstention reduces unsafe outputs, unanswered questions may limit usefulness, and the system does not guarantee completeness even for in-scope topics.

Second, our evaluation is based on a single real-world testing session (21 participants, 290 questions) and a single expert annotator. Larger-scale studies, multiple annotators, and additional clinical expert review would strengthen reliability estimates and enable inter-annotator agreement analysis. Performance may also vary across languages and modalities (text vs. voice), and further stratified analysis is needed.

Third, the system uses LLM components for routing and re-ranking, which may exhibit instability across model versions and can introduce biases. In addition, parts of the stack may depend on proprietary services (e.g., messaging platform APIs, LLM endpoints, or TTS), which can affect reproducibility and long-term maintainability.

Finally, we evaluate informational QA rather than EHR-based clinical decision support. The system should not be used to replace professional medical care, and any real deployment would require careful governance, monitoring, and periodic clinical review of content and behavior.

13. Bibliographical References

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

Christopher Baethge. 2008. The languages of medicine. *Deutsches Arzteblatt International*, 105(3):37.

Oleksandr Balabanov and Hampus Linander. 2025. Uncertainty quantification in fine-tuned llms using lora ensembles.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers.

Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. 2025. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI*, 4:e58670.

Abdullah Ashraf Hamad, Jaber H Jaradat, Hamza K Alsalhi, and Ibraheem M Alkhawaldeh. 2024. Medical research production in native languages: A descriptive analysis of pubmed database. *Qatar Medical Journal*, 2024(1):21.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams.

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.

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.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. Aradice: Benchmarks for dialectal and cultural capabilities in llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218.

Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. 2021. A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12).

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering.

Faisal Qarah and Tawfeeq Alsanoosy. 2025. Evaluation of arabic large language models on moroccan dialect. *Engineering, Technology & Applied Science Research*, 15(3):22478–22485.

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.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. 2023. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1).

Politimi Valkimadi, Drosos Karageorgopoulos, Harissios Vliagoftis, and Matthew Falagas. 2009. Increasing dominance of english in publications archived by pubmed. *Scientometrics*, 81(1):219–223.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pages 4442–4457.

Xinhua Zhu, Xuechen Yang, and Hongchao Chen. 2018. A biomedical question answering system based on snomed-ct. In *Knowledge Science, Engineering and Management*, pages 16–28, Cham. Springer International Publishing.

Evaluating the Retrieval Component in a Retrieval-Augmented Summarization System for Patient Records in French

Marco Naguib[1,2], **Christel Gérardin**[3,4], **Victor Beaucoté**[5,6], **Cyril Charron**[5], **Adrien Joseph**[5,7], **Aurélie Névéol**[1], **Xavier Tannier**[3]

[1] Université Paris-Saclay, CNRS, LISN, Orsay, France
[2] Innovation and Data Unit, IT Department, AP–HP, Paris, France
[3] LIMICS, Université Sorbonne Paris-Nord, Inserm, Sorbonne Université, Paris, France
[4] Service de médecine interne, Hôpital Tenon, AP–HP, Paris, France
[5] Medical Intensive Care Unit, Ambroise Paré Hospital, AP–HP, Boulogne-Billancourt, France
[6] UFR Simone-Veil Santé, UVSQ, Montigny-le-Bretonneux, France
[7] Inserm, CESP, Paris-Saclay University, UVSQ, Villejuif, France

Abstract

In emergency and intensive care settings, clinicians must process large volumes of patient data to make time-sensitive decisions. Summarizing patient records can help reduce cognitive load and improve decision-making, but the complexity and variability of clinical documentation create challenges. This study explores a Retrieval-Augmented Generation (RAG) approach, consisting of two phases: (1) retrieval of relevant clinical information, and (2) generation of a summary. This paper evaluates the retrieval component of RAG systems, focusing on its performance in clinical contexts. Using French clinical text, we assess retrieval models and propose an annotation-based querying method to improve accuracy and consistency in retrieving core clinical information. We use an annotated dataset from the AP–HP hospital to benchmark retrieval models tailored for French clinical records. The proposed annotation-based querying method is compared to traditional prompt-based approaches, demonstrating improved retrieval performance. The findings indicate that specialized retrieval techniques enhance the effectiveness of RAG systems in clinical settings, providing more accurate and relevant information for summarization. The study contributes to the development of clinical decision support tools by improving the retrieval process in RAG systems. The proposed methods offer a structured approach to summarizing patient records, which may help clinicians manage information more efficiently.
Keywords: Electronic Health Records, Retrieval-Augmented Generation (RAG), Information Retrieval

1. Introduction

In emergency and intensive care, clinicians must rapidly process extensive patient data for critical decisions. Efficient summarization of patient records can alleviate cognitive burdens and enhance decision-making by ensuring vital details such as patient allergies are accessible at the point of care.

Summarizing patient records is inherently challenging. Clinical data is complex, variable over time, and often inconsistently documented, which complicates the development of generalizable summarization systems: (Pivovarov and Elhadad, 2015; Keszthelyi et al., 2023). Additionally, traditional metrics like BLEU and ROUGE may not effectively assess the quality of summarization systems because they are designed to capture surface similarity with a reference summary rather than factual accuracy and relevance.

Recent advances in Retrieval-Augmented Generation (RAG) architectures offer promising solutions to these challenges by explicitly grounding generated summaries in relevant source documents. RAG combines aspects of extractive summarization—through retrieving relevant document passages—with abstractive methods by constructing coherent summaries from these retrieved passages. While RAG has shown potential in general-domain summarization tasks: (Lewis et al., 2020; Liu et al., 2024; Edge et al., 2025), applications to clinical summarization—particularly in languages other than English—remain underexplored: (Alkhalaf et al., 2024; Ji et al., 2024). Specifically, there is a notable lack of systematic evaluation of retrieval methods tailored to clinical texts in French, as well as limited exploration of effective querying strategies that leverage clinical annotations.

In this study, we address these gaps by systematically evaluating and improving the retrieval component of RAG architectures for patient record summarization in French clinical settings. We focus on enhancing retrieval effectiveness, leaving the detailed evaluation of the generation component for future work.

The main contributions of this study include:

1. **Benchmarking Retrieval Methods for French Clinical Text:** We establish a systematic evaluation framework and provide comprehensive benchmarks comparing sparse (e.g., TF-IDF) and dense retrieval methods (sentence embedding approaches)

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 57–65
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

on French clinical documents, addressing an evaluation gap in the literature.

2. **Annotated Clinical Dataset:** We annotate real-world clinical documents provided by the AP–HP. This benchmark enables rigorous evaluation of retrieval and summarization techniques in complex clinical settings.

3. **Novel Annotation-Based Querying Method:** We propose an innovative querying approach that leverages clinical annotations, consistently outperforming traditional prompting methods and significantly enhancing retrieval performance, especially for computationally efficient methods like TF-IDF.

By establishing a robust evaluation framework and demonstrating the effectiveness of annotation-driven retrieval, our work lays a strong foundation for future research on generating structured, clinically relevant summaries that can effectively support medical decision-making. We share the code of our experiments at https://github.com/marconaguib/rag.

2. Related Work

Patient record summarization is essential in medical informatics for condensing extensive clinical texts into concise summaries that aid healthcare professionals: (Alkhalaf et al., 2024; Ji et al., 2024). Early methods relied on extractive techniques, selecting key sentences to maintain factual accuracy but often resulting in incoherent summaries: (Mishra et al., 2014; Moen et al., 2016).

Advancements in natural language processing have shifted focus to abstractive summarization, which generates novel text for more fluent and coherent summaries: (Zhang et al., 2018, 2020). However, ensuring factual accuracy and handling specialized medical terminology remain significant challenges: (Kryściński et al., 2020; Maynez et al., 2020).

The Retrieval-Augmented Generation (RAG) framework has shown promise in improving factual accuracy by grounding generated content in relevant source documents (Lewis et al., 2020). In clinical settings, RAG enhances both the accuracy and relevance of summaries by leveraging patient data: (Alkhalaf et al., 2024; Uapadhyay and Viviani, 2025).

Evaluating the factual correctness of summaries is critical, as traditional metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are insufficient for clinical accuracy (Zhang et al., 2020). Specialized evaluation frameworks such as FRAMES have been developed to better assess retrieval and reasoning in RAG systems (Krishna et al., 2024). Additionally, methods like Facts as a Function (FaaF) and factuality optimization techniques using reinforcement learning have been proposed to mitigate inaccuracies in generated text: (Katranidis and Barany, 2024; Cai et al., 2024).

Recent work has also explored explanable evaluation paradigms for summarization quality beyond aggregate scores. Herserant and Guigue (2025) propose a framework that decomposes summary evaluation into atomic sentence-level assessments, offering improved interpretability and detailed alignment between source text and generated summaries, addressing a key limitation of conventional evaluation metrics that only provide holistic scores.

Our work builds on these developments by applying RAG-based patient record summarization to real-world clinical data, with a particular emphasis on retrieval component effectiveness and innovative annotation-based querying methods. Additionally, we contribute the first systematic benchmark comparing sentence embeddings specifically tailored to French clinical documents, complementing recent general-domain evaluation frameworks (Ciancone et al., 2024), and further informing model selection and retrieval strategies in clinical NLP.

3. Data and Annotation

Our study leverages a corpus of 10,000 clinical documents from the AP–HP hospital, approved by an internal ethics review board. For our experiments, we selected individuals with a median number of documents per patient (=61) to represent typical clinical scenarios realistically.

The clinical document corpus was annotated by four critical care physicians. Annotators were instructed to identify and label clinically relevant information within patient records, focusing on essential details required for structured clinical summaries. To quantify annotation consistency and reliability, three patients were sampled for inter-annotator agreement computation. Additionally, eight other patients were independently annotated by a single annotator to expand our dataset, totaling 11 patients and 671 documents.

Annotations were categorized into six predefined clinical labels, selected to comprehensively represent patient profiles and effectively support clinical decision-making:

Lifestyle: Lifestyle-related details such as living arrangements and daily activities that can impact health outcomes.

Medical history: Key elements of the broader medical history, including chronic disorders which may impact treatment options.

Surgical history: Details regarding the patients surgical history, providing context for current health status.

Treatments: Data on medications and therapies received, which are essential for understanding ongoing care.

Allergies: Information about the patients allergies is crucial for avoiding adverse effects.

Biometrics: Critical biometric measurements (e.g., weight, height) that inform clinical decisions.

The category definitions for clinical features were used as annotation guidelines.

4. Experiments

4.1. Retrieval-Augmented Generation Pipeline

Our Retrieval-Augmented Generation (RAG) pipeline is illustrated in Figure 1. It consists of two sequential steps—retrieval and generation—applied iteratively to construct a structured clinical summary from heterogeneous patient records.

Step 1: Retrieval Clinical documents are first segmented into fixed-size text chunks, which constitute the retrieval units. Then, for each target clinical section (e.g., medical history), relevant passages are retrieved from the patient record using either vocabulary-based methods (TF-IDF, BM25) or transformer-based embedding models. As shown in Figure 1, this step acts as a filtering mechanism that selects a small set of section-specific excerpts from a large and unstructured collection of clinical documents.

Step 2: Generation In the second step, a large language model (LLM) synthesizes the retrieved passages into a coherent and structured summary tailored to the clinical section of interest. Rather than processing the entire patient record, the LLM operates on the targeted subset produced by the retrieval step, which constrains the input and guides generation.

Although generation is an essential component of the overall pipeline, the analysis presented in this work focuses exclusively on Step 1, namely retrieval performance. Preliminary experiments conducted with the complete pipeline on our clinical data (to be reported in the supplementary materials in the final version) indicate that effective retrieval remains critical, even when using large-context LLMs. In particular, when provided with entire patient records instead of retrieved excerpts, the model tended to generate overly long and poorly structured summaries, whereas retrieval-based inputs led to more focused and clinically relevant outputs.

4.2. Retrieval Queries

We investigated seven query formulation strategies aimed at improving retrieval performance and, in turn, the quality of patient record summarization. These strategies include both manually defined textual queries and data-driven queries automatically derived from annotated clinical data.

Textual Queries. The first five strategies rely on explicit textual formulations expressed in natural language. They differ in how the target information is described and constrained, but all provide a text-based query to the retriever:

- **Specification-Based Prompts**: queries derived from clinician-defined specifications that explicitly describe the expected content of each summary section, closely mirroring clinical guidelines.
- **Examples**: queries composed of representative example phrases illustrating typical linguistic realizations of the targeted clinical information.
- **Keywords**: curated lists of salient terms and entities associated with each clinical section, enabling focused lexical or semantic matching.
- **Contextual Prompts**: queries that situate the requested information within its broader clinical context, encouraging retrieval of passages jointly mentioning related concepts, temporal cues, or recent assessments.
- **Structure Prompts**: queries that explicitly specify the expected organizational structure by enumerating fields or subcomponents to be retrieved.

Data-Driven Queries. The remaining two strategies construct query representations automatically from annotated clinical data, without relying on manually written text queries. Their underlying principle is illustrated in Figure 2.

- **Query-from-Annotations**: for a given clinical label, all annotated spans are encoded and averaged to form a single query vector that serves as a semantic prototype of the target concept.
- **Query-from-Annotation-Chunks**: a chunk-level variant in which all text chunks containing at least one annotation of the target label are encoded and averaged, capturing a broader contextual signal (Figure 2).

The two data-driven strategies are evaluated in a leave-one-out cross-validation (LOOCV) setting. For each experiment, query vectors evaluated on

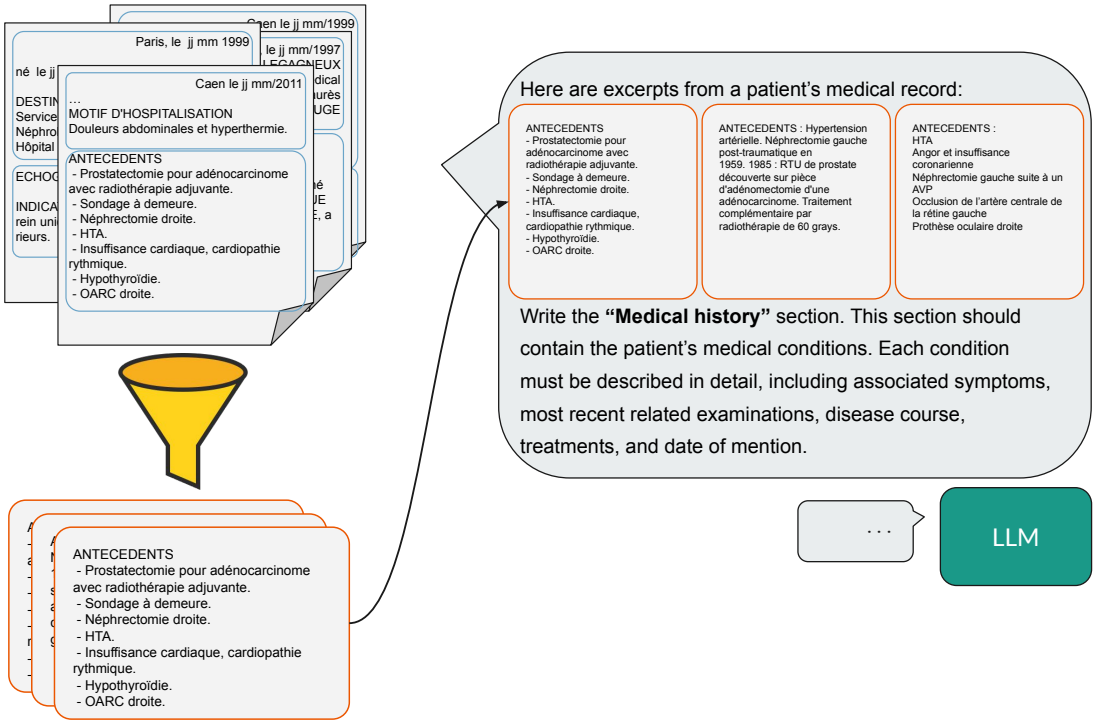


Figure 1: Overview of the Retrieval-Augmented Generation (RAG) pipeline for structured clinical summarization. Relevant passages are first retrieved from heterogeneous patient records for a given clinical section (e.g., medical history) and then provided to a large language model to generate a section-specific summary.

one patient record are constructed exclusively from annotations belonging to other patient records, ensuring that annotations from the test patient record are never used during query construction.

To evaluate the effectiveness of these query strategies, we benchmark them using a range of sparse and dense retrieval models, described next.

4.3. Retrieval Models

We evaluate seven retrieval models, selecting the top-performing open-source sentence embedding models for French as identified by the recent general-domain benchmark of Ciancone et al. (2024). Our evaluation complements this benchmark by systematically comparing these embeddings specifically on clinical texts.

TF-IDF: Traditional retrieval method based on term frequency-inverse document frequency.

BM25: Probabilistic retrieval method extending TF-IDF with term saturation and document length normalization. Note that BM25 is incompatible with our annotation-based querying approach, as it relies on explicit query terms rather than standalone query embeddings derived from annotations.

dangvantuan/sentence-camembert-large: CamemBERT-based sentence transformer optimized for French semantic retrieval.

BAAI/bge-m3: Multilingual embedding model from BAAI, optimized for diverse retrieval tasks.

OrdalieTech/Solon-embeddings-large-0.1: Large-scale embedding model designed for specialized-domain retrieval.

manu/sentence_croissant_alpha_v0.4: French embedding model emphasizing semantic similarity.

jinaai/jina-embeddings-v3: Multilingual embeddings optimized for efficient neural search.

4.4. Metrics

We evaluate the retrieval performance using recall at various reduction rates, specifically retrieving the top 5%, 10%, 15%, and 20% of chunks within each patient profile (i.e., the passage-level retrieval units described in Section 4.1). For each reduction rate, we compute the recall across all annotated labels to assess how effectively each retrieval model retrieves the most pertinent information.

In addition, we measure the computational efficiency of each model by recording the average computation time on an NVIDIA A100 GPU and on a 64-core CPU. This dual assessment helps us understand the trade-off between retrieval accuracy and computational demands, which is critical for practical deployment in clinical environments.

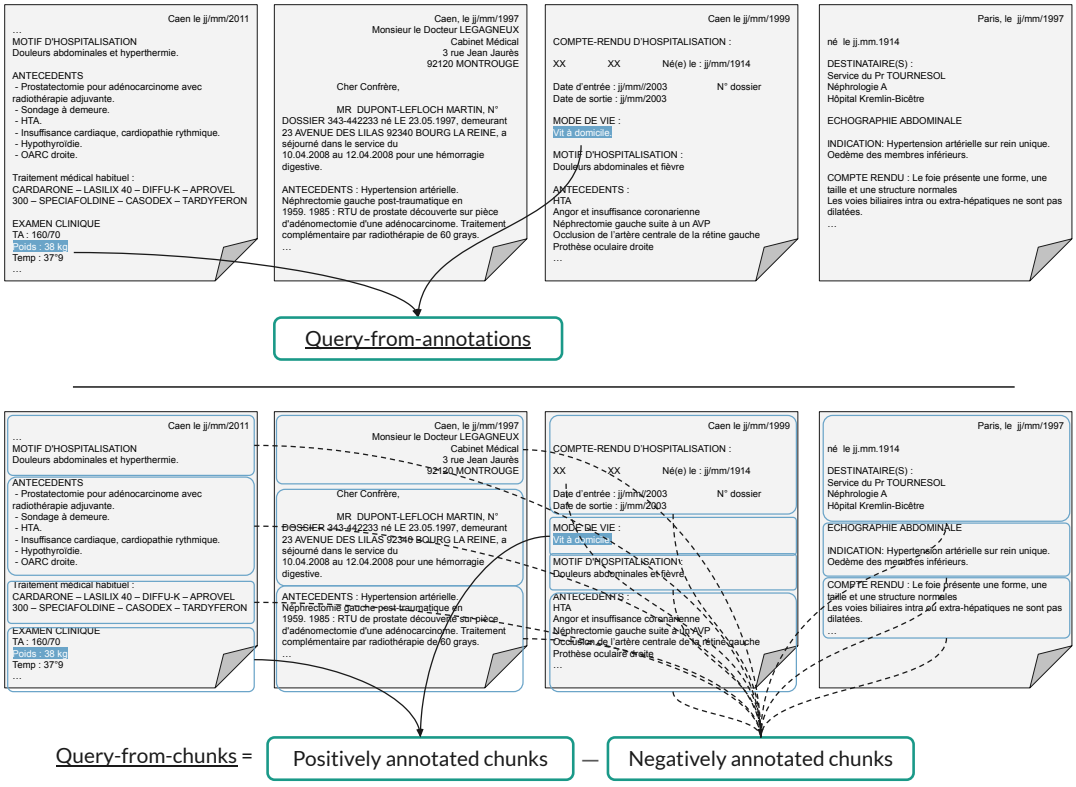


Figure 2: Illustration of data-driven query construction from annotations. Annotated spans or chunks associated with a given clinical label are extracted from multiple patient records and encoded into vector representations. In the *Query-from-Annotations* approach, only annotated spans are averaged to form a query vector, whereas *Query-from-Annotation-Chunks* averages representations of all text chunks containing at least one annotation. The resulting query vectors are then used to retrieve relevant passages in unseen patient records.

5. Results and Discussion

5.1. Inter-annotator Agreement

Inter-annotator agreement was measured to assess annotation consistency across clinical categories (Table 1). Overall agreement was moderate (average Cohen's $\kappa = 0.56$), reflecting both the intrinsic complexity of clinical records and subtle differences in annotation practices. Agreement varied across categories, with higher consistency for Lifestyle, Surgical History, and Allergies, and lower agreement for Medical History and Biometrics.

Pairwise analyses further highlight this variability, with κ values ranging from 0.43 to 0.86 depending on the entity type and annotator pair. Categories such as Lifestyle and Allergies exhibit consistently high agreement, whereas Medical History and Biometrics show greater divergence, underscoring the ambiguity of these clinical concepts.

To account for this variability, we adopted an additive consensus approach for the first three patients, aggregating annotations from all annotators when evaluating retrieval performance. Recall was computed against this merged reference, ensuring that the evaluation captures the full spectrum of expert judgment and mitigates the impact of individual annotation differences. This strategy provides a more robust and comprehensive benchmark for assessing retrieval and summarization performance under realistic clinical uncertainty.

Table 1: Inter-annotator agreement by category. Pairwise κ reported as min–max.

Category	Avg. κ	Median κ	Min–Max
Overall	**0.56**	**0.56**	0.51–0.63
Lifestyle	0.79	0.80	0.72–0.86
Medical history	0.52	0.51	0.45–0.62
Surgical history	0.67	0.67	0.61–0.73
Treatments	0.60	0.60	0.57–0.63
Allergies	0.66	0.66	0.49–0.81
Biometrics	0.52	0.52	0.43–0.66

5.2. Retrieval

Figure 3 shows the macro-average recall at 20% reduction for all query types and models. Using annotations as queries consistently improves retrieval performance compared to hard prompts. Notably, TF-IDF benefits significantly from annotation-based

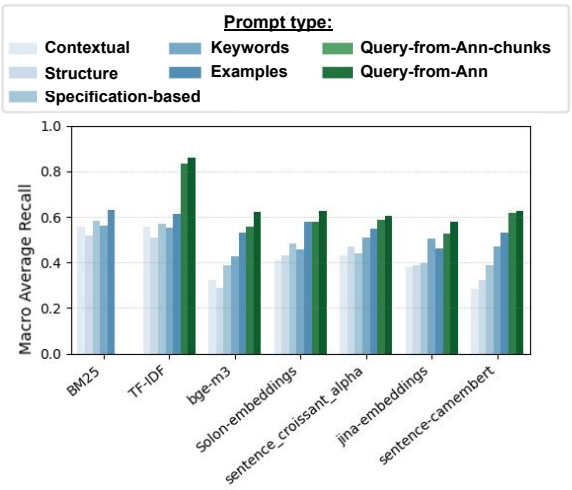


Figure 3: Macro-average Recall at 20% reduction per retrieval model and query type. Annotation-based queries (green) consistently improve retrieval performance.

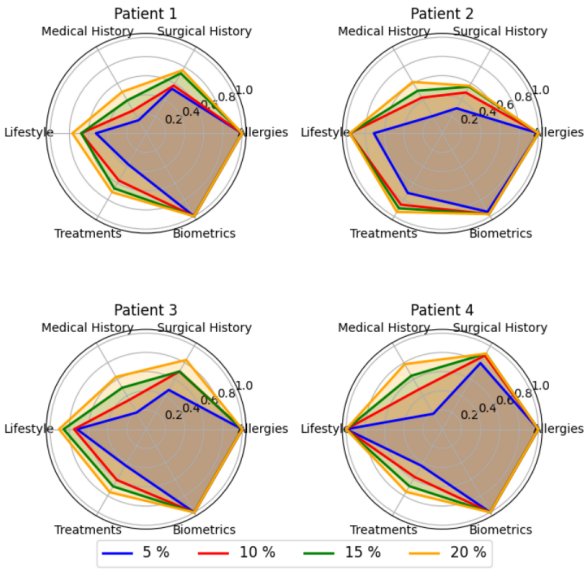


Figure 4: Recall across clinical sections for individual patient profiles at different reduction rates.

querying, surpassing transformer-based models in recall and being about 100 times faster when a GPU is available, while also being CPU-friendly (Table 2). Given its computational efficiency and strong retrieval performance, TF-IDF is particularly suitable for clinical settings, where speed and reliability are critical.

Finally, Figure 4 illustrates retrieval recall across different annotation categories for four patient profiles, highlighting variability in retrieval difficulty across clinical information types. These profiles include the three patients evaluated using the Additive consensus approach (1, 2, and 3). As anticipated by clinicians, performance is higher for

Table 2: Average computation times of encoding all patient profile chunks on an NVIDIA A100 GPU and on a 64-core CPU.

Model	NVIDIA A100 (s)	64-core CPU (s)
sentence-camembert	4.88	160
bge-m3	4.76	170
Solon-embeddings	4.71	173
sentence-croissant-alpha	19.4	716
jina-embeddings	2.18	161
TF-IDF	-	0.05
BM25	-	0.03

compact categories such as *allergies* compared to *medical history*, where information is scattered across documents and changes more over time.

6. Conclusion

This work investigated the retrieval component of Retrieval-Augmented Generation (RAG) systems for patient record summarization in French clinical settings. Through a systematic evaluation of seven retrieval models and multiple query formulation strategies, we demonstrated that retrieval effectiveness can be substantially improved by leveraging clinical annotations to construct query representations. Across models and reduction rates, annotation-based queries consistently outperformed manually designed textual prompts, highlighting the importance of data-driven querying strategies for clinical information retrieval.

A key finding of our study is that simple and computationally efficient methods, such as TF-IDF, can achieve strong retrieval performance when coupled with annotation-derived queries—often matching or surpassing transformer-based embedding models at a fraction of the computational cost. This result is particularly relevant for real-world clinical deployment, where robustness, interpretability, and resource efficiency are critical constraints. Our benchmark further provides the first systematic comparison of sentence embedding models on French clinical documents, complementing existing general-domain evaluations with clinically grounded insights.

By isolating and rigorously evaluating the retrieval stage, this work establishes a solid foundation for the development and assessment of retrieval-augmented clinical summarization systems. Future work will extend this evaluation to the generation component, with a focus on factual accuracy, clinical relevance, and end-user utility, as well as on reducing reliance on manual annotations to improve scalability.

7. Acknowledgments

This work has received funding from the PARTAGES project, awardee of the Bpifrance France 2030 call for proposals “Digital Commons for Generative Artificial Intelligence.”

Limitations

Our evaluation primarily focuses on the retrieval step of the proposed retrieval-augmented summarization pipeline. While retrieval quality is a critical component, it does not necessarily translate directly into the overall performance of the full system, particularly in terms of summarization accuracy and clinical usability. Future work will aim to assess the impact of retrieval effectiveness on downstream tasks, such as information synthesis and generation, to provide a more holistic evaluation of the pipeline.

Furthermore, our evaluation concentrated on retrieval performance and computational efficiency but did not comprehensively address other important factors such as potential biases in retrieval outcomes or the interpretability of retrieved results. These aspects remain open areas for further investigation, especially given the clinical context in which fairness, transparency, and robustness are paramount.

The annotated corpus used in this study also presents certain limitations. First, due to confidentiality agreements, we are unable to publicly share the dataset or the corresponding annotations, restricting reproducibility and external validation. While this is common in clinical research, it underscores the broader challenge of benchmarking patient record summarization models on real-world data. Second, the number of patient records in the corpus was relatively small, reflecting the pilot nature of this study. Despite this, the annotations demonstrated moderate to substantial agreement, suggesting that expert consensus was reasonably achieved. The annotation process involved 671 documents, highlighting the complexity and effort required in manual Electronic Health Record (EHR) summarization.

Additionally, the current methodology relies heavily on high-quality clinical annotations to construct effective query representations. While this allows for precise and structured retrieval, it also limits scalability and generalizability, as such annotations are time-consuming and costly to obtain in broader settings. Exploring strategies for reducing annotation dependence—such as semi-supervised methods or automated labeling—will be an important direction for future work.

Future work will seek to address these limitations by scaling the dataset, refining evaluation metrics, and incorporating a broader range of factors—including potential biases and real-world clinical applicability—into our assessment framework.

Ethical Considerations

This study involves the use of de-identified patient data approved (CSE 21-36) by an internal ethics review board at the AP–HP hospital. Throughout our work, we made every effort to maintain patient privacy and confidentiality in compliance with ethical and institutional guidelines.

The application of recommendation or automated summarization systems in clinical care carries potential risks, such as automation bias, where clinicians may overly trust system outputs without sufficient scrutiny. Such biases can substantially impact patient outcomes if incorrect or incomplete information goes unnoticed.

Our findings specifically highlight the value of simpler and more interpretable methods, such as TF-IDF retrieval, which not only match but can outperform sophisticated sentence embedding models when coupled with our query-from-annotation method. Besides offering improved interpretability, TF-IDF is computationally lighter, thus aligning better with environmentally responsible and sustainable AI practices.

We recognize the critical implications of integrating automated tools into clinical workflows. Therefore, we emphasize the necessity of extensive evaluation beyond retrieval performance. Our future research steps include a comprehensive assessment of the full retrieval-augmented generation (RAG) pipeline, addressing not only factual accuracy and summary quality but also clinical relevance, user trust, overall usability, and eventual impacts on clinical decisions and patient outcomes.

8. Bibliographical References

Mohammad Alkhalaf, Ping Yu, Mengyang Yin, and Chao Deng. 2024. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of Biomedical Informatics*, 156:104662.

Tianchi Cai, Zhiwen Tan, Xierui Song, Tao Sun, Jiyan Jiang, Yunqi Xu, Yinger Zhang, and Jinjie Gu. 2024. FoRAG: Factuality-optimized Retrieval Augmented Generation for Web-enhanced Long-form Question Answering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 199–210, New York, NY, USA. Association for Computing Machinery.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. MTEB-French: Resources for French Sentence Embedding Evaluation and Analysis. *arXiv preprint arXiv:2405.20468*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From Local to Global: A Graph RAG Approach to Query-Focused Summarization.

Tanguy Herserant and Vincent Guigue. 2025. Seval-ex : Un paradigme basé sur les phrases atomiques pour une évaluation explicable de la qualité des résumés. In *Actes de la 20e Conférence en Recherche d'Information et Applications (CORIA)*, pages 217–229, Marseille, France. ATALA \\& ARIA.

Yuelyu Ji, Zhuochun Li, Rui Meng, Sonish Sivarajkumar, Yanshan Wang, Zeshui Yu, Hui Ji, Yushui Han, Hanyu Zeng, and Daqing He. 2024. RAG-RLRC-LaySum at BioLaySumm: Integrating Retrieval-Augmented Generation and Readability Control for Layman Summarization of Biomedical Texts.

Vasileios Katranidis and Gabor Barany. 2024. FaaF: facts as a function for the evaluation of RAG systems. *arXiv e-prints*, pages arXiv–2403.

Daniel Keszthelyi, Christophe Gaudet-Blavignac, Mina Bjelogrlic, Christian Lovis, et al. 2023. Patient information summarization in clinical settings: scoping review. *JMIR Medical Informatics*, 11(1):e44639.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2024. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *arXiv preprint arXiv:2409.12941*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9332–9346.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

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.

Shengjie Liu, Jing Wu, Jingyuan Bao, Wenyi Wang, Naira Hovakimyan, and Christopher G Healey. 2024. Towards a Robust Retrieval-Based Summarization System.

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.

Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R. Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol. 2014. Text summarization in the biomedical domain: A systematic review of recent research. *Journal of Biomedical Informatics*, 52:457–467. Special Section: Methods in Clinical Research Informatics.

Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. 2016. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37.

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.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Rishabh Uapadhyay and Marco Viviani. 2025. Enhancing Health Information Retrieval with RAG by Prioritizing Topical Relevance and Factual Accuracy. *arXiv preprint arXiv:2502.04666*.

Yuhao Zhang, Qian Chi, and Xibin Xia. 2018. Learning to Summarize Radiology Findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.

Team MKC at MEDIQA-SYNUR 2026: Retrieval-Augmented Generation Based Nurse Observation Extraction

Kyomin Hwang[1], Nojun Kwak[1,2†]
[1]GSCST, Seoul National University, [2]AIIS, Seoul National University
{kyomin98, nojunk}@snu.ac.kr

Abstract

Recent advancements in Large Language Models (LLMs) have played a significant role in reducing human workload across various domains, a trend that is increasingly extending into the medical field. In this paper, we propose an automated pipeline designed to alleviate the burden on nurses by automatically extracting clinical observations from nurse dictations. To ensure accurate extraction, we introduce a method based on Retrieval-Augmented Generation (RAG). Our approach demonstrates effective performance, achieving an F1-score of 0.796 on the MEDIQA-SYNUR test dataset.

Keywords: Clinical Observation Parsing, Retrieval-Augmented Generation

1. Introduction

Recent advancements in Large Language Models (LLMs), exemplified by the GPT series (Brown et al., 2020), have accelerated a paradigm shift across a multitude of domains, demonstrating remarkable capabilities in natural language understanding and generation. This momentum has profoundly impacted the medical field (Kim et al., 2024a), leading to the development of specialized models such as MedGemma (Sellergren et al., 2025) and LLaVA-Med (Li et al., 2023a). These models have achieved state-of-the-art performance on various biomedical tasks by leveraging large-scale, domain-specific datasets. However, obtaining such high-quality annotated data remains a significant bottleneck due to privacy concerns and the high cost of expert annotation. In this context, the SYNUR (Corbeil et al., 2025) dataset has established a crucial benchmark, facilitating the automatic extraction of clinical observations from nursing dictations—a task essential for streamlining clinical documentation.

Building upon these developments, we propose a novel Retrieval-Augmented Generation (RAG) pipeline designed to accurately extract clinical observations from complex nurse dictations. Unlike conventional approaches that rely on resource-intensive fine-tuning, our automated pipeline integrates a synergistic dual-retrieval mechanism. First, we employ an ontology-based retrieval system to fetch the most relevant medical concepts based on the current utterance, ensuring terminological precision. Second, we utilize a memory bank to retrieve semantically similar dialogue segments and their corresponding gold-standard observations, enabling the model to learn extraction patterns via in-context learning. This dual approach equips the foundation LLM with both explicit medical context and structural guidance. Utilizing this training-free framework, we achieved an F1-score of 0.796 on the MEDIQA-SYNUR test set. These results demonstrate that our approach can effectively alleviate the nursing documentation burden by automating the interpretation of patient interactions with high reliability.

† Corresponding author.

2. Related Works

2.1. Foundation model in Medical Domain

Recent advancements in large-scale data acquisition and the exponential growth of computing power have led to the emergence of foundation models capable of performing diverse tasks within a unified framework (Jang et al., 2023; Brown et al., 2020). This paradigm has significantly influenced the medical domain, triggering the development of specialized foundation models. Early efforts focused on encoder-based models like ClinicalBERT (Huang et al., 2019) and BlueBERT (Peng et al., 2019), which are optimized for medical knowledge representation and retrieval. More recently, generative and multi-modal models such as LLaVA-Med (Li et al., 2023a) and MedGemma (Sellergren et al., 2025) have been introduced to handle complex biomedical reasoning. Currently, active research is dedicated to adapting these generalist foundation models to enhance their performance on specific clinical tasks (Li et al., 2023b; Toma et al., 2023). Aligning with this research process, this paper proposes a method for developing a Large Language Model (LLM) specifically optimized for analyzing medical observations within nurse dictations.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 66–72
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

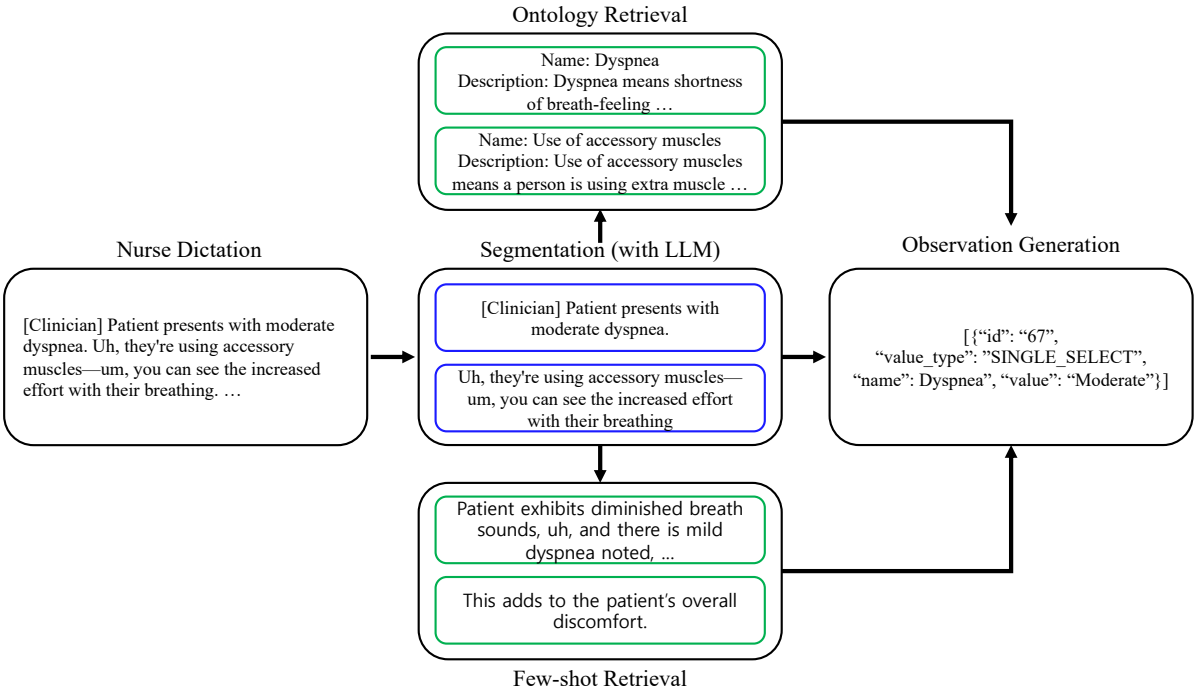


Figure 1: Full illustration of Retrieval-Augmented Generation (RAG) Based Nurse Observation Extraction Pipeline

2.2. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Kim et al., 2024b) has emerged as a prominent approach for tailoring foundation models to specific tasks. While foundation models possess vast general knowledge, they often struggle with domain-specific nuances and factual accuracy in critical fields like healthcare. RAG addresses this limitation by grounding the model's responses in retrieved evidence. Previous studies have demonstrated that utilizing RAG techniques can significantly enhance task-specific performance across various domains without the need for additional training of the foundation models (Chen et al., 2022; Kim et al., 2024b). Leveraging this capability to handle complex clinical terminologies and contexts, we propose a RAG-based automatic generation pipeline designed to specialize a foundation LLM for extracting medical observations from nurse dictations, all without requiring further model training.

3. Method

In this section, we present a pipeline designed to automatically extract clinical observations from nurse dictations. Figure 1 illustrates our medical observation pipeline.

3.1. Nurse Dictation Segmentation

Nurse dictations processed by LLMs are typically lengthy. Therefore, following the approach proposed by (Corbeil et al., 2025), we first segment the input dictation into distinct units containing clinical facts. The prompt used for this segmentation is shown in Figure 2.

Each nurse dictation is divided into multiple segments. For each segment, we generate observations by providing the LLM with two types of retrieved context: 1) relevant schemas retrieved from the ontology, and 2) semantically similar segments retrieved from the training dataset. The latter serves as few-shot examples, consisting of segment-observation pairs, to guide the LLM in generating the correct observation for the input segment.

3.2. Description for Medical Observation in Ontology

First, we recognized that directly inputting medical observations defined within an ontology (e.g., weight-bearing status) into an LLM could lead to performance degradation. Consequently, we implemented a process to append descriptions to each medical observation. Figure 3 illustrates the prompt used to generate these descriptions.

3.3. Retrieving Relevant Schemas from Ontology

To enhance the accuracy of observation extraction, we identify schemas relevant to the input segment from the ontology and provide them as input to the

Prompt for Nurse Dictation Segmentation

You are an expert clinical documentation assistant.

Given an input TRANSCRIPT of a nurse's observations about a patient, divide the TRANSCRIPT into contiguous SEGMENTS based on clinical facts.

A clinical fact refers to specific, verifiable information related to the health of a patient (symptoms, vitals, exam findings, interventions, response to treatment, labs/imaging results, functional status, safety events, etc.).

SEGMENTATION RULES:
- Output segments MUST preserve the original order.
- Segments MUST be contiguous chunks of the original transcript (do NOT reorder).
- Do NOT paraphrase; keep the segment text as close to verbatim as possible.
- Include all information from the transcript (no omissions).
- Split when the clinical topic/fact changes (new symptom, new vital sign, new intervention, new assessment finding, new timeframe, etc.).
- If multiple sentences describe one tightly-related clinical fact, keep them in the same segment.
- If the transcript contains speaker tags (e.g., [Nurse], [Patient]) keep them as they appear.

OUTPUT FORMAT:
Return ONLY a valid JSON object with exactly one key:
{
"segments": [
"<segment 1>",
"<segment 2>",
...
]
}

No additional keys. No extra text outside JSON.

TRANSCRIPT:

{NURSE DICTATION}

Figure 2: Prompt for nurse dictation segmentation

Description generation prompt

You are a helpful and professional medical assistant.
Your task is to explain the given medical assessment item (Name)
so that a patient or a layperson can easily understand it.
Provide the explanation using simple and friendly vocabulary.
Keep the explanation concise (1-2 sentences) and strictly avoid technical jargon.
Explain the meaning of '{term}'

Figure 3: Prompt for description generation

LLM. For retrieval, we employ a hybrid approach utilizing both BlueBert (Peng et al., 2019) and TF-IDF. We format each schema using its name, description and its value enumerations (options). The input format is as follows:

name. Description Options: Option1, Option2 (1)

We select the top 10 most relevant schemas to include in the LLM input.

3.4. Retrieving Few-Shot Examples from Training Dataset

In addition to schemas, we utilize few-shot examples to improve generation performance by retrieving segments from the training dataset that are similar to the input segment, along with their ground-truth observations.

To construct the retrieval pool, we first processed the training dataset using ChatGPT. We extracted individual segments and their corresponding observations to create a database of pairs. Figure 4 illustrates the prompt used for this extraction process,

Prompt for Segment Generation for Train dataset

You are a precise medical documentation assistant.
Your task is to segment the transcript strictly by **sentences** and map observations to them.

CLINICAL CONCEPT DEFINITIONS
Use the following definitions to understand the context of the observations. This will help you match the transcript text to the correct observation even if the wording is slightly different.
{ONTOLOGY LIST}

SEGMENTATION RULES:
1. **Unit:** The minimum segment unit is a full grammatical **sentence**. Do not split a single sentence into smaller phrases or clauses, even if it contains multiple facts.
2. **Completeness:** The entire transcript must be reconstructed in the "segments" list sequentially, including [Clinician].

MAPPING RULES:
1. **Multiple Matches:** Since a single sentence can contain multiple clinical facts, the "observations" list for a segment can contain multiple items.
2. **Relevance:** Include an observation in a segment ONLY if that specific sentence explicitly mentions or supports that observation (refer to the Definitions above).
3. **Empty List:** If a sentence is just conversational filler (e.g., "Hello," "Okay") and contains no specific medical orders/observations from the list, return an empty list `[]` for "observations".

CRITICAL CONSTRAINTS (MUST FOLLOW):
1. **NO OMISSIONS:** Every single object in the provided "Observations" input list **MUST** be mapped to at least one segment.
2. **Check Before Output:** Before generating the final JSON, verify that the count of unique observation IDs in your output matches the count of observation IDs in the input. If an observation is missing, find the most relevant sentence and map it there.

OUTPUT FORMAT:
Return a JSON object with a key "segments".
"segments": [
{{
"segment": "Full sentence text here.",
"observations": [{{ ...obs1... }}, {{ ...obs2... }}]
}},
...
]

Transcript:
{TRANSCRIPT}

Observations:
{OBSERVATIONS}

Figure 4: Prompt for segment generation for train dataset

Segment For Train Dataset
Segment: [Clinician] Patient presents with moderate dyspnea.
Observations: [{"id": "67", "value_type": "SINGLE_SELECT", "name": "Dyspnea", "value": "Moderate"}]
Segment: We're keeping an eye on all these symptoms to manage the patient's care effectively.
Observations: []

Figure 5: Segment Example for Train Dataset.

while Figure 5 displays examples of the resulting segment-observation pairs. From this processed dataset, we retrieve the top 15 examples to serve as few-shot context for the LLM. This retrieval step is performed using a hybrid method combining BlueBert and BM25.

3.5. Observation Generation

We retrieved N and K entries, respectively, from the ontology pool and the few-shot example pool generated in the previous section, instructing the LLM to extract medical observations relevant to the current utterance. Figure 6 presents the prompt used for this task.

4. Experiments

In this section, we present the evaluation results of our proposed approach and analyze the performance of various model configurations.

Prompt for Observation Generation

```
You are an expert clinical documentation assistant.
Your task is to extract clinical observations from the transcript segment and map them strictly to the provided SCHEMA.

### INSTRUCTIONS:
1. Review the TRANSCRIPT SEGMENT.
2. Extract ONLY information that matches the 'name' and definitions in the SCHEMA.
3. If no information matches the schema, return an empty list [].
4. For 'SINGLE_SELECT' or 'MULTI_SELECT', values MUST be from 'value_enum'.
5. Return ONLY a valid JSON list.

### RELEVANT SCHEMA:
{schema_str}

### REFERENCE EXAMPLES (JSONL format):
Below are similar examples from the training data. Follow the mapping logic shown in "observations".
{few_shot_text}

### CURRENT TRANSCRIPT SEGMENT:\n{transcript_segment}

Return ONLY the JSON list of observations.
```

Figure 6: Prompt for observation generation

Table 1: Performance comparison of the proposed methods. Best results are marked in **bold**.

Method	Precision	Recall	F1
+ Few-shot Ex	0.789	0.694	0.739
+ Few-shot Ex + Schema	**0.812**	**0.847**	**0.829**

4.1. Evaluation on Development Dataset

4.1.1. Overall Result

Table 1 summarizes the performance of different methods on the development dataset. In this experiment, we utilized GPT-5-mini as the backbone model. As shown in the table, incorporating schemas alongside few-shot examples yields superior performance in extracting automatic medical observations from nurse dictations, compared to providing few-shot examples alone.

4.1.2. Ablation on Retrieval Model

We conducted an ablation study on the retrieval models used to fetch schemas and few-shot examples. Table 3 summarizes the performance results. As shown in the table, BlueBERT achieves the best performance in terms of F1-score.

4.2. Evaluation on Test Dataset

In this section, we present the evaluation results on the MEDIQA-SYNUR test dataset, summarized in Table 2. Regarding model configuration, the combination of GPT-5.1 as the observation generator and GPT-5-mini as the segmentor yielded the best performance. We also observed a positive correlation between the number of few-shot examples and overall performance. Interestingly, while larger generator models generally improved performance (e.g., GPT-5.1 outperformed GPT-5-mini), this trend was not strictly monotonic. Specifically, GPT-5.2 exhibited performance degradation compared to GPT-5.1. Furthermore, the superior performance of the GPT-5-mini segmentor can be attributed to model consistency. Since the memory bank of segments was constructed using GPT-5-mini, employing the identical model for segmentation during inference ensures better alignment with the retrieved examples, leading to optimal results.

5. Discussion

Efficiency in Observation Extraction Pipeline In this study, we utilized in-context learning by retrieving medical terms from few-shot examples and an ontology. However, during the inference phase, when dictation is divided into segments, not all segments necessarily contain meaningful medical observations. Although this aspect was outside the scope of our current work, we propose that the pipeline's efficiency could be significantly improved by incorporating a preliminary classifier. This classifier would determine the presence of medical observations within a segment before prompting the LLM, thereby reducing unnecessary computational costs for irrelevant segments.

Table 2: Performance comparison with different configurations on the Test Dataset. †denotes the usage of refined prompt.

Base Model	Shots	LLM Backbone		Metrics		
		Generator	Segmenter	Precision	Recall	F1
	3	GPT-5-mini	GPT-5-mini	0.739	0.810	0.773
	5	GPT-5-mini	GPT-5-mini	0.744	0.817	0.779
	10	GPT-5-mini	GPT-5-mini	0.744	0.818	0.779
BlueBert	10	GPT-5.1	GPT-5-mini	0.785	0.805	0.795
	10	GPT-5.2	GPT-5-mini	0.766	0.825	0.794
	10	GPT-5.2	GPT-5.2	0.759	**0.827**	0.792
	15	GPT-5.1	GPT-5-mini	**0.786**	0.807	**0.796**

Table 3: Ablation on retrieval model.

Method	Precision	Recall	F1
ClinicalBert	0.811	0.845	0.828
BlueBert	0.807	0.860	0.833
PubMedBert	0.805	0.842	0.823

Dataset Analysis Our analysis indicates opportunities to improve data consistency. For example, values such as 37.5°C sometimes appear as "375", and unit annotations (*e.g.,* Temperature Unit) are not always consistent. Standardizing value formats and units would likely improve training efficiency and retrieval accuracy.

6. Conclusion

In this paper, we proposed a Retrieval-Augmented Generation (RAG)-based pipeline designed to automatically extract medical observations from nurse dictations. To achieve this, we constructed a memory bank leveraging both a medical observation ontology and previously annotated observation tags from existing dictations. By integrating these dataset into the LLM generation process, our approach guides the model to produce more accurate outputs. Consequently, our method demonstrated its effectiveness by achieving an F1-score of 0.796 on the MEDIQA-SYNUR test dataset.

7. Acknowledgements

This work was supported by the Korean Government through the grants from IITP (RS-2021-II211343, RS-2022-II220320, RS-2025-25442338).

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.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 859–870, Suzhou (China). Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Jiho Jang, Chaerin Kong, DongHyeon Jeon, Seonhoon Kim, and Nojun Kwak. 2023. Unifying vision-language representation space with single-tower transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):980–988.

Hyeonjin Kim, Min Kim, Jae Jang, KiYoon Yoo, and Nojun Kwak. 2024a. TEAM MIPAL at MEDIQA-

M3G 2024: Large VQA models for dermatological diagnosis. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 334–338, Mexico City, Mexico. Association for Computational Linguistics.

Taehoon Kim, Pyunghwan Ahn, Sangyun Kim, Sihaeng Lee, Mark Marsden, Alessandra Sala, Seung Hwan Kim, Bohyung Han, Kyoung Mu Lee, Honglak Lee, et al. 2024b. Nice: Cvpr 2023 challenge on zero-shot image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7356–7365.

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.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

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*.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

Automatic Generation of Discharge Summaries Using Large Language Models: A Systematic Literature Review

Lucas Molino Piñar, Manuel Carlos Díaz Galiano, María Teresa Martín Valdivia
SINAI Research Group, Universidad de Jaén, Spain
{lmolino, mcdiaz, maite}@ujaen.es

Abstract

Discharge summaries are critical documents for continuity of care, yet their manual creation imposes significant burdens on clinical staff. This systematic literature review examines current approaches to automatic generation of discharge summaries using Natural Language Processing (NLP) and Large Language Models (LLMs). Following the Kitchenham guidelines for systematic reviews in software engineering, we searched Scopus and PubMed databases for studies published between 2023 and 2026, identifying 9 primary studies from an initial pool of 102 papers. Our analysis reveals that GPT-4 and its variants dominate current research (appearing in 6 of 9 studies), while open-source alternatives like LLaMA show promise for privacy-preserving deployments. Evaluation primarily relies on automatic metrics (ROUGE, BLEU) combined with human expert assessment. Key challenges include hallucination rates ranging from 33% to 64%, information omission, integration with Electronic Health Record (EHR) systems, and context window limitations. Studies addressing factuality employ human-in-the-loop validation, prompt engineering techniques, and knowledge graph-based correction mechanisms. Despite these challenges, recent implementations demonstrate clinical feasibility, with one study achieving a 94.35% System Usability Score. This review provides a comprehensive synthesis of the state-of-the-art and identifies opportunities for future research in this rapidly evolving field.

Keywords: Large Language Models, Discharge Summaries, Natural Language Processing, Clinical Documentation, Systematic Review

1. Introduction

Discharge summaries are essential clinical documents that communicate critical patient information between healthcare providers, ensuring continuity of care after hospitalization (Kripalani et al., 2007). These documents typically include diagnoses, treatments administered, hospital course, and discharge recommendations. However, their manual creation represents a significant burden for clinicians, contributing to documentation fatigue and reducing time available for direct patient care (Sinsky et al., 2016).

The emergence of Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023), and their clinical adaptations has created unprecedented opportunities for automating clinical documentation tasks. These transformer-based architectures have demonstrated remarkable capabilities in text generation, summarization, and domain adaptation, making them promising candidates for addressing the discharge summary automation challenge.

1.1. Motivation and Scope

The motivation for this systematic review stems from three key observations. First, the clinical documentation burden continues to escalate, with physicians spending up to 50% of their time on documentation tasks. Second, recent advances in LLM technology (2023-2026) have dramatically improved the feasibility of automatic text generation in specialized domains. Third, there is a critical need to understand the current state of research, including both achievements and limitations, to guide future development efforts in clinical NLP.

This review focuses specifically on the automatic generation of discharge summaries intended for healthcare professionals (physician-to-physician communication). We explicitly exclude studies focusing on patient-friendly simplification of existing summaries, as this represents a distinct NLP task with different objectives. Our temporal scope (2023-2026) captures the most recent developments in LLM-based clinical text generation, particularly following the release of GPT-4 and similar advanced models.

1.2. Objectives and Research Questions

The primary objective of this systematic literature review is to identify, analyze, and synthesize current approaches for automatic generation of discharge summaries using NLP and LLMs. Specifically, we aim to catalog the NLP models and techniques employed for discharge summary generation, identify evaluation methodologies and metrics used to assess generated summary quality, document technical and software engineering challenges encountered during implementation, and examine how issues of factual accuracy, hallucinations, and patient

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 73–82
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

data privacy are addressed.

To guide this systematic review, we formulated four research questions addressing complementary aspects of automatic discharge summary generation. **RQ1** asks what language models and NLP techniques are currently used for automatic generation of discharge summaries. **RQ2** investigates what evaluation metrics are employed to validate the clinical and linguistic quality of generated reports. **RQ3** explores the main technical and software engineering challenges reported in implementing these systems. Finally, **RQ4** examines how issues of factuality, hallucinations, and medical data privacy are addressed in these systems. These questions were designed to cover the complete lifecycle of discharge summary generation systems, from model selection (RQ1), through quality assessment (RQ2), to practical implementation considerations (RQ3 and RQ4).

The remainder of this paper is organized as follows: Section 2 describes the systematic review protocol, including the search strategy and selection criteria. Section 3 presents the results organized by research question, analyzing models, evaluation metrics, technical challenges, and factuality issues. Section 4 discusses our findings and their implications for practice. Finally, Section 5 concludes with limitations and future work directions.

2. Review Protocol

This systematic review follows the guidelines proposed by Kitchenham and Charters (2007) for conducting systematic literature reviews in software engineering. The review was conducted using the Parsifal tool[1] for protocol management and study tracking. Data extraction sheets and quality assessment scores are available from the authors upon request.

2.1. PICOC Framework

We defined the scope of our review using the PICOC framework (Petticrew and Roberts, 2006), an acronym for Population, Intervention, Comparison, Outcome, and Context. The **Population** comprises software systems for automatic generation of clinical documentation, specifically discharge summaries. The **Intervention** encompasses NLP techniques, large language models (LLMs such as GPT, BERT, LLaMA), natural language generation methods, sequence-to-sequence architectures, and transformer-based approaches. For **Comparison**, we considered traditional template-based methods, manual generation by clinicians, rule-based systems, different model architectures (transformer vs. RNN), and extractive vs. abstractive approaches. The expected **Outcome** includes quality of generated reports (linguistic and clinical), evaluation metrics (ROUGE, BLEU, BERTScore, human assessment), technical feasibility, implementation viability, and clinical acceptance. The **Context** covers hospital environments, electronic health record (EHR) systems, clinical documentation workflows, medical software engineering, and healthcare informatics.

2.2. Search Strategy

We selected two complementary databases for our search: **Scopus** as the primary source for its comprehensive coverage of computer science and medical informatics literature, and **PubMed/MEDLINE** as a secondary source for its specialized coverage of biomedical literature. We initially considered IEEE Xplore (3 results) and ACM Digital Library (14 results, only 1-2 within temporal scope), but excluded them due to insufficient coverage of our topic. ACL Anthology and related computational linguistics venues (EMNLP, BioNLP, LOUHI workshops) were not included as separate sources; however, Scopus indexes proceedings from these venues, ensuring their coverage in our search.

Table 1 presents the keywords derived from the research questions and their synonyms used to construct the search string.

Table 1: Keywords and synonyms used in search strategy.

Keyword	Synonyms
discharge summary	discharge report, hospital summary, clinical summary, discharge letter, discharge note
automatic generation	automated generation, AI generation, machine-generated, computer-generated
large language model	LLM, GPT, BERT, language model, neural language model, transformer
natural language processing	NLP, text processing, computational linguistics, automatic summarization
deep learning	machine learning, artificial intelligence, neural networks

The search string was constructed using Boolean operators following the structure: `(keyword1 OR synonym1) AND (keyword2 OR synonym2)`. The following search string was applied to Scopus on January 6, 2026:

```
TITLE-ABS-KEY(("discharge summary" OR "discharge report" OR
```

[1] https://parsif.al

```
"hospital summary" OR "clini-
cal summary") AND ("automatic
generation" OR "automated gen-
eration" OR "AI generation" OR
"LLM" OR "large language model"
OR "natural language generation"
OR "text generation") AND ("NLP"
OR "natural language processing"
OR "deep learning" OR "machine
learning" OR "artificial intel-
ligence")) AND PUBYEAR > 2022
AND (LIMIT-TO(DOCTYPE, "ar") OR
LIMIT-TO(DOCTYPE, "cp"))
```

This search returned 94 papers. The equivalent PubMed query, executed on January 7, 2026, was:

```
("discharge sum-
mary"[Title/Abstract] OR "dis-
charge report"[Title/Abstract]
OR "clinical sum-
mary"[Title/Abstract])
AND ("large language
model"[Title/Abstract] OR
"LLM"[Title/Abstract] OR
"natural language genera-
tion"[Title/Abstract] OR
"GPT"[Title/Abstract]) AND
("2023"[Date - Publication] :
"2026"[Date - Publication])
```

This search returned 8 papers. The total initial pool was 102 papers.

2.3. Selection Criteria

Studies were included if they met all of the following criteria: published between 2023 and 2026, written in English, focused on discharge summaries for healthcare professionals, presented specific methods for automatic discharge summary generation, included empirical evaluation with real or synthetic clinical data, appeared in a peer-reviewed venue or high-quality preprint server with subsequent peer-reviewed publication, constituted a full paper (minimum 6 pages, not extended abstract), and addressed NLP, AI, or software engineering aspects of the problem.

Conversely, studies were excluded if they were secondary studies (systematic reviews, literature reviews, surveys), tertiary studies (reviews of reviews), abstract-only or limited access without full text, only mentioned discharge summaries tangentially, had unclear or absent methodology, represented grey literature without peer review, were duplicate publications, addressed the wrong domain (not discharge summaries), or focused on patient-friendly simplification only.

2.4. Quality Assessment

Each candidate study was evaluated using six quality assessment criteria, scored on a 3-point scale of 0 (no), 0.5 (partially), or 1 (yes). The criteria assessed whether results are valid and conclusions supported by evidence (QA1), whether research objectives and questions are clearly defined (QA2), whether methodology is described in sufficient detail to allow replication (QA3), whether experiments or evaluations are reported using real or synthetic data (QA4), whether study limitations and threats to validity are discussed (QA5), and whether the contribution is relevant to software engineering or clinical NLP (QA6).

Studies scoring ≥ 3.5 out of 6 were included in the final analysis. Additional rigor criteria applied: sample size $n < 10$ was considered insufficient; non-blinded evaluation without justification was penalized; purely fictitious data without real-world validation resulted in rejection.

2.5. Data Extraction

For each included study, we extracted 29 data fields organized into five categories. **Bibliographic** data included venue and publication type. **Study Context** (addressing RQ1) covered research goal, domain, dataset, models, architecture, training approach, and input/output formats. **Evaluation** data (addressing RQ2) encompassed metrics, metric values, human evaluation details, baseline comparisons, and best results. **Software Engineering** aspects (addressing RQ3) included technical challenges, system architecture, EHR integration, computational requirements, and software tools. Finally, **Factuality and Ethics** data (addressing RQ4) covered hallucination handling, privacy/security measures, clinical validation, and limitations.

2.6. Study Selection Process

Table 2 illustrates our selection process following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Page et al., 2021).

Table 2: PRISMA flow: study selection.

Phase	N	Excluded
Initial search	102	–
Duplicates removed	101	1
Relevance screening	20	81 (off-topic)
Title/abstract	14	6 (no focus)
Full-text + QA	**9**	5 (see text)

From the initial 102 papers, we removed 1 duplicate. During relevance screening, 81 papers were excluded for being off-topic (e.g., general clinical NLP without discharge focus), addressing wrong domains (radiology reports, pathology), or focusing on patient-friendly simplification. Title/abstract

screening excluded 6 additional papers: 3 lacked specific generation methods, 2 were surveys or position papers, and 1 was an extended abstract. During quality assessment, five papers were rejected: two used completely fictitious data without real-world validation, one had sample size $n = 10$ deemed insufficient, one focused exclusively on patient simplification, and one had methodological issues (single non-blinded evaluator). One included study (Williams et al., 2025) was initially available as a medRxiv preprint but has since been accepted for peer-reviewed publication in PLOS Digital Health (June 2025); we retained it given its methodological rigor and subsequent peer review.

3. Results

Table 3 summarizes the 9 primary studies included in this review. It is important to note that while eight studies utilized real clinical records, Grazhdanski et al. (2025) employed entirely synthetic data generated from MSD Manuals. This distinction should be carefully considered when interpreting cross-study comparisons, particularly regarding faithfulness and hallucination metrics.

3.1. RQ1: Models, Architectures, and Training Approaches

Our analysis reveals a clear hierarchy of model preferences in discharge summary generation. GPT-4 and its variants dominate current research, appearing in 6 of 9 studies (67%). Table 4 presents detailed performance comparisons across different models.

Williams et al. (2025) provided the most direct comparison between GPT-4 and GPT-3.5-turbo on identical emergency department data. GPT-4 achieved significantly better results across all error categories: 33% completely error-free summaries versus only 10% for GPT-3.5-turbo, with hallucination rates of 42% versus 64% respectively. GPT-4 also produced more concise outputs (median 235 words vs. 369.5 words) with better readability scores (Flesch-Kincaid Grade Level 10.0 vs. 10.7).

LLaMA 3 appears in three studies, offering a privacy-preserving alternative to commercial APIs. Hains et al. (2025) demonstrated that locally-deployed LLaMA 3 models achieved 61.6-61.9% of human performance using a validated 17-domain scoring tool (Savvopoulos et al.). Notably, the 70B parameter model showed no statistically significant improvement over the 8B model ($p = 0.974$), suggesting diminishing returns at higher parameter counts for this task. Li et al. (2025) compared base and fine-tuned LLaMA 3 variants, finding that fine-tuning with 102 clinical cases produced summaries with length closest to the gold standard, but did not consistently outperform zero-shot GPT-4 on quality metrics. This suggests that prompt engineering may be more cost-effective than fine-tuning for current applications.

It is important to note that this field evolves rapidly; as newer and more capable models emerge (such as GPT-5 or future LLaMA iterations), the model preferences documented here will likely shift accordingly. The general architectural patterns and evaluation methodologies, however, are expected to remain relevant.

Three distinct architectural paradigms emerge from our analysis. The most common approach (5 studies) relies on single-model API-based generation using commercial APIs (GPT-4, GPT-4o), which offers rapid prototyping but raises privacy concerns and limits reproducibility. A second paradigm involves multi-agent orchestration: Croxford et al. (2025) evaluated Microsoft MagenticOne, a multi-agent framework with specialized agents for planning, web search, file handling, and coding, while Trejo Omeñaca et al. (2025) implemented custom multi-agent orchestration with iterative prompt engineering for real-world deployment. The third paradigm, explored by Fu et al. (2025), proposes a hybrid structured-unstructured pipeline combining extractive summarization (TextRank with ClinicalBERT embeddings) for unstructured notes with data-to-text generation for structured EHR data (medications, labs).

Regarding training approaches, the overwhelming majority of studies (7/9) employed zero-shot prompting without model fine-tuning. Schwieger et al. (2024) conducted extensive prompt engineering over 6 weeks with hundreds of iterations, highlighting that prompt optimization is the current preferred approach over fine-tuning. Only Li et al. (2025) performed supervised fine-tuning on clinical data, and Grazhdanski et al. (2025) employed a self-refinement approach based on Madaan et al.'s method. The preference for zero-shot approaches reflects both the strong out-of-box capabilities of modern LLMs and the challenges of acquiring sufficient clinical training data.

3.2. RQ2: Evaluation Paradigms and Metrics

Our analysis reveals a significant paradigm shift in evaluation methodology. Early approaches relied solely on automatic metrics (ROUGE, BLEU), but contemporary studies recognize that clinical quality cannot be captured by lexical overlap metrics alone. Eight of nine studies (89%) now combine automatic metrics with human expert evaluation, reflecting a maturing understanding of the multidimensional nature of clinical text quality.

Table 5 presents the specific metric values re-

Table 3: Summary of included primary studies.

Study	Venue	Domain	Models	N	QA
Williams et al. (Williams et al., 2025)	PLOS Digit. Health	Emergency	GPT-4, GPT-3.5	100	5.0
Croxford et al. (Croxford et al., 2025)	npj Digit. Med.	General	GPT-4o, LLaMA	200	6.0
Schwieger et al. (Schwieger et al., 2024)	Int. J. Med. Inf.	Psychiatry	GPT-4	20	6.0
Challener et al. (Challener et al., 2025)	J. Hosp. Med.	Multi-spec.	GPT-4o (Epic)	100	5.0
Trejo et al. (Trejo Omeñaca et al., 2025)	Computers	Multi-spec.	Multiple LLMs	47	4.5
Hains et al. (Hains et al., 2025)	Intern. Med. J.	General	LLaMA 3	10	4.0
Li et al. (Li et al., 2025)	J. Biomed. Inf.	Oncology	GPT-4, LLaMA 3	50	5.0
Grazhdanski et al.	J. Biomed. Inf.	Synth. Gen.	GPT-based, KG	900	4.5
Fu et al. (Fu et al., 2025)	IEEE BigData 2024	General	BART, PEGASUS	MIMIC	4.0

ported across studies that employed automatic evaluation metrics. Not all studies reported comparable automatic metrics; some relied primarily on human evaluation or used domain-specific instruments.

ROUGE metrics appear in 7 of 9 studies, but reported values are notably low (ROUGE-L: 0.16-0.24). This reflects the abstractive nature of discharge summary generation, good summaries need not share exact phrasing with source documents. Fu et al. (2025) demonstrated that supplementing unstructured summaries with structured data improved ROUGE-L by 5.6 percentage points (from 17.6% to 23.2% for the Lead-3 baseline). Semantic similarity metrics appear as alternatives to lexical overlap: Li et al. (2025) reported semantic similarity scores of 0.83-0.837, with LLaMA 3 achieving the highest similarity to reference summaries despite lower ROUGE scores.

Human evaluation approaches vary significantly in rigor and standardization. Table 6 summarizes the methodologies across studies.

The PDSQI-9 instrument (Provider Documentation Summarization Quality Instrument) emerges as the most rigorously validated tool. Croxford et al. (2025) demonstrated that it achieves excellent psychometric properties (ICC 0.867) and captures nine clinically relevant dimensions: Cited, Accurate, Thorough, Useful, Organized, Comprehensible, Succinct, Synthesized, and Stigmatizing. Schwieger et al. (2024) developed a comprehensive 15-item rubric specifically for psychiatric summaries, evaluating dimensions including text coherence ($\Delta 0.96$ human advantage), summarization quality ($\Delta 1.07$), and specificity ($\Delta 1.19$). Importantly, AI showed advantages in conciseness (low unnecessary information $\Delta 0.07$) and safety/legal aspects ($\Delta 0.01$).

A notable methodological development is the LLM-as-Judge paradigm, which has gained traction in NLP evaluation since 2023 (Zheng et al., 2023). In the context of clinical summarization, Croxford et al. (2025) systematically evaluated this approach for discharge summary assessment. This paradigm uses LLMs to automatically evaluate clinical summaries, potentially scaling human-quality assessment. Key findings from their systematic evaluation reveal that reasoning models excel, with GPT-o3-mini achieving ICC 0.818 (95% CI: 0.772-0.854), the highest correlation with human experts. DeepSeek R1 achieved ICC 0.762, demonstrating that open-source reasoning models can approach commercial performance. Non-reasoning models underperform: GPT-4o achieved only ICC 0.730, and the multi-agent framework MagenticOne achieved ICC 0.768. The speed advantage is substantial, with LLM evaluation completing in 22 seconds versus 10 minutes for human experts, a 27$\times$ speedup. Reasoning models particularly excelled in attributes requiring deeper analysis (Cited, Organized, Synthesized, Thorough), while simpler attributes showed less differentiation. This paradigm offers a promising path toward scalable, reproducible evaluation of clinical text generation systems.

Two studies assessed readability using established metrics. Williams et al. (2025) found that GPT-4 achieved Flesch-Kincaid Grade Level 10.0 (IQR 9.5-11.1) and Flesch Reading Ease 48.6 (IQR 41.0-52.0), indicating 10th-grade reading level. Challener et al. (2025) reported that human summaries showed higher Flesch Reading Ease (33.11 vs 26.2, $p < 0.05$), indicating simpler language, though LLM summaries scored higher on clinical utility.

Table 4: Detailed model performance comparison across studies.

Model	Type	Study	Score	Notes
GPT-4	Comm. API	Williams et al.	33% err-free	Halluc. 42% vs 64% in 3.5
GPT-4	Comm. API	Schwieger	Mean 3.12/5	40% halluc.; 17% "ready"
GPT-4	Comm. API	Li et al.	Rel. 4.95/5	Factuality 4.40/5
GPT-4o	Epic Integ.	Challener et al.	All superior	Superior all 9 PDSQI-9
GPT-4o	Comm. API	Li et al.	Compl. 4.55/5	Best ROUGE-L (0.16)
GPT-o3-mini	Reasoning	Croxford et al.	ICC 0.818	Best inter-rater reliability
LLaMA 3 8B	Open Source	Hains et al.	19.1/31	Locally deployable
LLaMA 3 70B	Open Source	Hains et al.	19.2/31	No sig. diff vs 8B
Fine-tuned	Open Source	Li et al.	BLEU 0.04	Best length alignment
PEGASUS	Pre-trained	Fu et al.	+5.6% R-L	Best w/ struct. data

Table 5: Automatic metric values reported in studies using quantitative evaluation.

Study	R-1	R-2	R-L	BLEU	Other Metrics
Li et al. (GPT-4o)	0.38	0.11	0.16	–	Sem. Sim.: 0.83
Li et al. (LLaMA)	–	–	0.16	0.04	Sem. Sim.: 0.837
Fu et al. (base)	–	–	0.176	–	–
Fu et al. (+struct)	–	–	0.232	–	+5.6% improvement
Trejo et al.	0.43	0.22	0.24	–	Spanish (es-es), n=35

Table 6: Human evaluation methodologies across studies.

Study	Raters	Blind	Instrument
Schwieger et al.	8	Yes	Custom 15-item
Challener et al.	5	Yes	Mod. PDQI-9
Williams et al.	2	No	Error taxonomy
Croxford et al.	Expert	Yes	PDSQI-9
Grazhdanski et al.	2	Yes	SUS (10 item)
Trejo-Omeñaca et al.	Clinical	No	4-pt Usefulness

3.3. RQ3: Technical Challenges and System Architectures

Hallucinations and information omission represent the most critical barriers to clinical deployment. Table 7 presents detailed error analysis across studies.

Williams et al. (2025) developed a detailed hallucination taxonomy identifying four primary error patterns: redacted information filling (where LLMs inferred content from redacted/de-identified portions), fabricated follow-up (hallucinated specialty appointments not arranged), incorrect precautions (fabricated ED return instructions), and plan conflation (interim plans reported as final discharge plans). Notably, hallucinations were most concentrated in the Plan and Recommendations sections, which paradoxically are most critical for patient safety.

Information omission emerged as equally problematic: Williams et al. (2025) reported 47% omission rates, while Trejo Omeñaca et al. (2025) found 53% of summaries missing critical information, particularly social context (52%) and temporal details. This suggests that LLMs may prioritize fluency over completeness when generating clinical text.

The 0% hallucination rate reported by Challener et al. (2025) warrants careful interpretation in this context: Epic's tool uses a proprietary, highly restrictive prompt that excludes structured data (labs, medications, procedures) and existing summaries, limiting the scope for factual errors. Additionally, the study employed blinded physician reviewers using a validated 9-item instrument (PDSQI-9), which may have different sensitivity thresholds than the error taxonomies used in other studies. This methodological heterogeneity makes direct comparisons across studies challenging.

Context window constraints force significant compromises across implementations. Schwieger et al. (2024) limited inputs to 1,000-5,000 tokens for GPT-4's 8,192 context window, reserving space for output generation. Challener et al. (2025) reported that Epic's tool requires multi-pass processing when chart data exceeds context limits, with each iter-

Table 7: Detailed hallucination and error analysis across studies.

Study	Hall.	Inacc.	Omiss.	Error Types
Williams et al.	42%	10%	47%	Hallucinated follow-up, redacted info filled, incorrect precautions
Schwieger et al.	40%	30%	–	37.5% relevant; concentrated in recommendations
Trejo et al.	47%	–	53%	Medication dose (56%), social context (52%), temporal inconsistency
Hains et al.	Low	1 case	–	Inc. COVID antiviral; patient/clinician name confusion
Challener et al.	0%[†]	–	–	Restrictive prompts; excludes structured data

ation building on previous outputs. Hains et al. (2025) excluded ICU notes and certain documentation types entirely due to naming convention differences. Trejo Omeñaca et al. (2025) reported lost-in-the-middle effects for lengthy documents, where information in the middle of context windows is poorly utilized.

Two production deployment architectures were evaluated in detail. The Epic Systems Integration described by Challener et al. (2025) uses GPT-4o with a proprietary Epic prompt, automatic extraction of prioritized clinical notes, multi-pass processing for oversized charts, and excludes structured data (labs, procedures) and existing summaries. However, its black-box prompt design limits reproducibility. The Custom Hospital Integration implemented by Trejo Omeñaca et al. (2025) at Hospital General de Granollers, Spain, employs multi-agent orchestration with specialized agents, provides bilingual support (Spanish/Catalan), and followed a three-phase development process: proof-of-concept, prototype, and pilot. During the pilot phase, 47 discharge reports were generated with physicians rating them 2.9/4 average usefulness.

Regarding computational requirements, studies predominantly rely on API-based access, limiting computational transparency. API-based approaches (GPT-4, GPT-4o) dominate due to ease of integration. Local deployment, as demonstrated by Hains et al. (2025), enables privacy preservation but requires local GPU infrastructure. Evaluation with LLM-as-Judge completed in 22 seconds per summary (Croxford et al., 2025).

3.4. RQ4: Factuality, Privacy, and Toward Autonomous Systems

All production-deployed systems currently require physician review. However, our analysis identifies three strategies that may eventually reduce human-in-the-loop dependency. First, knowledge graph validation as proposed by Grazhdanski et al. (2025) enables automated correction using SPARQL queries against medical ontologies (SNOMED, LOINC, RxNorm, DrugBank); notably, their 93.65% faithfulness score was achieved on *synthetic* data generated from MSD Manuals, not real clinical records, and may not generalize to production settings. Second, the LLM-as-Judge framework described by Croxford et al. (2025) demonstrates that reasoning models (GPT-o3-mini, ICC 0.818) can reliably flag problematic summaries for human review, enabling selective rather than universal physician oversight. Third, structured data integration as shown by Fu et al. (2025) reduces hallucination potential for factual content by supplementing free-text generation with verified structured EHR data (medications, labs, procedures).

Table 8 presents comprehensive quality metrics across studies.

Three strategies address privacy concerns in clinical text generation. Local model deployment, as demonstrated by Hains et al. (2025), shows that locally-deployed LLaMA 3 models avoid transmitting patient data to external servers, achieving GDPR/HIPAA compliance by design. Synthetic data training, as employed by Grazhdanski et al. (2025), generated 900 synthetic discharge summaries grounded in MSD Manuals without any real patient data, enabling model training and benchmarking without privacy risks. De-identification preprocessing, used by multiple studies (Williams et al., Schwieger et al.), involves applying pseudonymization before API submission, though this introduces risk of residual identifiers.

Based on our analysis, we identify conditional scenarios where reduced human oversight may become acceptable: low-risk sections (automated generation of administrative content such as demographics and medication lists from structured sources with minimal oversight), high-confidence outputs (selective review triggered by LLM-as-Judge quality scores below threshold, e.g., ICC < 0.8), knowledge graph verification (factual claims verified against medical ontologies before inclusion), and dual-LLM verification (independent generation and cross-validation using separate model instances). However, current hallucination rates (33-47%) and the concentration of errors in safety-critical sections (Plan, Recommendations) indicate that full human review remains essential for patient safety.

Table 8: Quality scores achieved across studies.

Study: Key Quality Score	Study: Key Quality Score
Croxford et al.: ICC 0.818 (LLM-as-Judge)	Hains et al.: 19.2/31 (61.9% of human)
Grazhdanski et al.: SUS 94.35%, Faithfulness 93.65%	Trejo-Omeñaca et al.: 2.9/4 usefulness
Challener et al.: LLM superior all 9 PDSQI-9 criteria	Williams et al.: 33% error-free (GPT-4)
Li et al.: Relevance 4.95/5, Completeness 4.55/5	Schwieger et al.: Human 3.78/5 vs AI 3.12/5

4. Discussion

Our review captures a critical inflection point in clinical NLP. Early MIMIC-based work (Johnson et al., 2016) established foundational datasets, while pre-transformer approaches struggled with long clinical documents. The emergence of GPT-3.5/GPT-4 in 2022-2023 marked the first clinically viable LLM capabilities for discharge summary generation.

This review reveals a rapidly maturing field where LLM-based discharge summary generation has progressed from proof-of-concept to production. GPT-4 dominates current research (6/9 studies), though open-source alternatives (LLaMA 3) address privacy concerns. The high hallucination rates (33-64%) represent the most significant barrier to autonomous deployment, though human summaries also contain errors (10%). Information omission rates up to 53% suggest LLMs prioritize fluency over completeness. The evaluation landscape is evolving toward validated instruments (PDSQI-9) and LLM-as-Judge approaches (Croxford et al., 2025).

Regarding study limitations, sample sizes ranged from 10 to 900, with most studies (6/9) using fewer than 100 cases. Only two studies (Croxford et al., 2025; Schwieger et al., 2024) employed blinded evaluation with multiple raters. All studies except one were single-center, and only Trejo Omeñaca et al. (2025) addressed multilingual settings. These factors suggest caution when generalizing findings.

In terms of practical implications, hybrid workflows are essential: LLM-generated summaries should serve as drafts requiring physician review. Furthermore, model selection must be carefully aligned with the clinical context: commercial APIs (e.g., GPT-4o) offer superior reasoning and text generation quality suitable for environments where infrastructure permits, whereas locally deployed open-source models (e.g., LLaMA) provide strict data privacy and regulatory compliance for sensitive EHR integrations. Regarding standardization, we propose minimum evaluation standards for future studies: clinical text generation evaluations must incorporate concurrent human expert assessment using validated instruments (such as the PDSQI-9) alongside automatic metrics, and transparently report inter-rater reliability. EHR-integrated solutions built upon these foundations demonstrate the most viable path to clinical utility.

5. Conclusions

This systematic literature review analyzed 9 primary studies on automatic generation of hospital discharge summaries using LLMs, published between 2023 and 2026. Our findings demonstrate that GPT-4 and its variants currently dominate research, appearing in 67% of studies, while open-source LLaMA alternatives offer privacy-preserving deployment options. Evaluation practices are maturing, with hybrid approaches combining automatic metrics (ROUGE, BLEU) with validated clinical instruments (PDQI, PDSQI-9) and emerging LLM-as-Judge paradigms. Hallucinations remain a critical challenge, with rates ranging from 33% to 64%, necessitating human-in-the-loop validation for clinical deployment. Privacy can be addressed through local model deployment, knowledge graph-based validation, and synthetic data generation methodologies.

Future research should focus on multi-center validation studies, specialty-specific prompt optimization, temporal consistency mechanisms to prevent propagation of outdated information, and robust automated hallucination detection systems. The field is poised for broader clinical adoption, pending resolution of safety and reliability concerns.

6. Limitations

This review has several limitations. We searched Scopus and PubMed; while Scopus indexes ACL/EMNLP proceedings, direct searching of ACL Anthology might have identified additional work. We included only English-language studies. Our temporal scope (2023-2026) excludes earlier foundational work, though we provided historical context. Quality assessment was performed by a single reviewer; multiple reviewers with inter-rater reliability (Cohen's κ) would strengthen validity. Publication bias may overrepresent positive results, and some reported metrics (0% hallucination, 93.65% faithfulness) may reflect methodological choices rather than generalizable performance. Furthermore, a formal meta-analysis or quantitative synthesis was not attempted because the extreme

heterogeneity of the selected studies, in terms of evaluation methodologies, mixed automatic and human evaluation instruments, and underlying data sources, precludes direct, reliable numerical comparison.

7. Ethical Considerations

The deployment of LLMs for clinical documentation raises significant ethical concerns. Patient safety remains paramount: hallucination rates of 33-64% documented in this review underscore that autonomous generation without human oversight is currently unacceptable. All production systems reviewed require physician verification before clinical use.

Privacy and data protection present ongoing challenges. Studies using commercial APIs (GPT-4) transmit patient data to external servers. Local deployment of open-source models (Hains et al., 2025) offers a privacy-preserving alternative, though with performance trade-offs. De-identification preprocessing mitigates but does not eliminate re-identification risks.

Transparency and accountability require that generated summaries be clearly marked as AI-assisted, allowing receiving clinicians to apply appropriate scrutiny. The "black-box" nature of commercial LLMs complicates error attribution and quality assurance.

Regarding research ethics, all included studies either obtained IRB approval, used publicly available de-identified data (MIMIC), or employed synthetic data. Future research should ensure appropriate ethical oversight and consider potential biases in training data that may perpetuate health disparities.

8. 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

9. Bibliographical References

Douglas Challener, Shant Ayanian, Alexander Ryu, John O'Horo, and Heather Heaton. 2025. Quality assessment of artificial intelligence-generated versus human-written hospital summaries evaluating detail, usefulness, and continuity of care. *Journal of Hospital Medicine*.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, Matthew M. Churpek, Anoop Mayampurath, Frank Liao, Cherodeep Goswami, Karen K. Wong, Brian W. Patterson, and Majid Afshar. 2025. Evaluating clinical ai summaries with large language models as judges. *npj Digital Medicine*, 8(1).

Jiaojiao Fu, Bowen Yang, Yi Guo, Yangfan Zhou, and Xin Wang. 2025. Automated clinical summary generation via integrating structured and unstructured data. *Communications in Computer and Information Science*, 2301:290 – 304.

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*, 170.

Lewis Hains, Oliver Kleinig, Ashwin Murugappa, Samuel Gluck, Jarrod Marks, Toby Gilbert, and Stephen Bacchi. 2025. Large language model discharge summary preparation using real-world electronic medical record data shows promise. *Internal Medicine Journal*, 55(7):1188 – 1192.

Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. Technical report, Keele University and Durham University.

Sunil Kripalani, Frank LeFevre, Christopher O. Phillips, Mark V. Williams, Preetha Basaviah, and David W. Baker. 2007. Deficits in communication and information transfer between hospital-based and primary care physicians: Implications for patient safety and continuity of care. *JAMA*, 297(8):831–841.

Yiming Li, Fang Li, Na Hong, Manqi Li, Kirk Roberts, Licong Cui, Cui Tao, and Hua Xu. 2025. A comparative study of recent large language models on generating hospital discharge summaries for lung cancer patients. *Journal of Biomedical Informatics*, 168.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

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, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjorn Hrobjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71.

Mark Petticrew and Helen Roberts. 2006. *Systematic Reviews in the Social Sciences: A Practical Guide*. Blackwell Publishing, Oxford, UK.

Arne Schwieger, Katrin Angst, Mateo de Bardeci, Achim Burrer, Flurin Cathomas, Stefano Ferrea, Franziska Grätz, Marius Knorr, Golo Kronenberg, Tobias Spiller, David Troi, Erich Seifritz, Samantha Weber, and Sebastian Olbrich. 2024. Large language models can support generation of standardized discharge summaries – a retrospective study utilizing chatgpt-4 and electronic health records. *International Journal of Medical Informatics*, 192.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Alex Trejo Omeñaca, Esteve Llargués Rocabruna, Jonny Sloan, Michelle Catta-Preta, Jan Ferrer i Picó, Julio Cesar Alfaro Alvarez, Toni Alonso Solis, Eloy Lloveras Gil, Xavier Serrano Vinaixa, Daniela Velasquez Villegas, Ramon Romeu Garcia, Carles Rubies Feijoo, Josep Maria Monguet i Fierro, and Beatriu Bayes Genis. 2025. Leave as fast as you can: Using generative ai to automate and accelerate hospital discharge reports. *Computers*, 14(6).

Christopher Y.K. 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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

10. Language Resource References

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.

Smart_solutions at MEDIQA-SYNUR 2026: A Multi-Stage LLM Pipeline for Nursing Observation Extraction

Prateek Munjal
M42
Abu Dhabi, UAE
pmunjal@m42.ae

Abstract

Extracting clinical observations from nursing dictations addresses an important problem of addressing burden in clinical documentation. In this work, we describe our approach submitted to MEDIQA-SYNUR 2026, which achieved third place among participating teams with a balanced precision and recall of 0.80 on the unseen test set. Our approach, instead of finetuning LLMs, is to adopt a multi-stage pipeline of agents: Observation Agent, Ontology Matching Agent, Relevance Scoring Agent, Evidence Assignment Agent, and Formatting Agent. First, the Observation Agent extracts clinical observations and corresponding evidence from the nurse transcript. These observations are then processed by the Ontology Matching Agent, which maps them to a restricted set of candidate ontology fields via TF-IDF–based retrieval, and subsequently evaluated by the Relevance Scoring Agent, which assigns continuous support scores (1–5) to each candidate field. Finally, field value assignments are performed by the Evidence-Based Agent, which extracts values strictly from nurse transcripts and clinical observations (Observation Agent outputs) to populate each ontology field. These outputs are then formatted by the Formatting Agent to ensure correct submission structure with the necessary metadata. Our agentic system results suggest that combination of agents with prompt engineering can narrow the gap between general and specialized clinical NLP models, making it an immediately deployable alternative to traditional fine-tuning.

Keywords: clinical information extraction, EHR structuring, LLMs, agents, clinical NLP, nursing transcripts

1. Introduction

Clinical nursing documentation (Willard-Grace et al., 2019; Friganović et al., 2019) is a crucial problem which helps in capturing continuous patient treatment and clinical decision-making. The nurses record patient observations frequently during the patient's stay in the hospital. These observations are typically mental status, functional ability, vital signs, and risk assessments, which are converted into structured EHR records. This process in practice is time-consuming, leading to a documentation burden for nurses(Padilla Fortunatti and Palmeiro-Silva, 2017; McHugh et al., 2011). It is important to note that automated extraction of nursing observations presents unique challenges compared to other clinical information extraction tasks. Nursing dictations are often conversational, fragmented, and context-dependent, while the target output is a large, fine-grained ontology of structured fields with strict value constraints. High-stakes domains like healthcare make the errors in this setting costly, highlighting the need for ideal system to prioritize both precision and recall metrics (Hurd et al., 2024; Altermatt et al., 2025).

The MEDIQA-SYNUR shared task (George Michalopoulos, 2026) addresses this problem by benchmarking systems for extracting observation from nurse dictations/transcripts, requiring the participants to map free-text nurse dictations to 192 structured EHR fields. These fields span binary, categorical, and free-text values. Unlike prior medical order or entity extraction tasks, MEDIQA-SYNUR task emphasizes on grounding and balanced precision–recall performance by measuring F1. Recent works (Ntinopoulos et al., 2025) in clinical NLP has increasingly explored LLMs for information extraction, often relying on single-stage, end-to-end prompting or domain specific fine-tuning of generalist models. While the finetuning approaches (Liu et al., 2025a; Majdik et al., 2024; Singhal et al., 2025; Christophe et al., 2024; Tu et al., 2024) demonstrate report strong performances gains, in this study we deliberately adopt to use the generalist LLM (Yang et al., 2025; Liu et al., 2024; Team et al., 2025; Grattafiori et al., 2024; Achiam et al., 2023) without any task specific training. We make this choice to benchmark immediately deployable baseline for resource-constrained clinical settings where fine-tuning is either infeasible or too costly.

In this paper, we present our approach of agentic system submitted to MEDIQA-SYNUR 2026, which rather than treating observation extraction as a single inference task, we decompose it into a multi-stage pipeline consisting of 5 agents: namely;(i) Observation Agent, (ii) Ontology Matching Agent, (iii) Relevance Scoring Agent, (iv) Evidence Assignment Agent, and (v) Formatting Agent. Our system achieves third place among participating teams on the MEDIQA-SYNUR test set, with balanced precision and recall of 0.80. The high F1 score

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 83–91
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

of 0.80 demonstrates that Agentic AI (Shimgekar et al., 2025) can support in automating extracting observations from nurse dictations, and help with documentation burden experienced by nurses.

2. Related Work

Clinical information extraction has traditionally been approached through supervised learning using domain-specific encoders and sequence labeling models. Early work (Zhang and Chen, 2022; Hu et al., 2024) focused on named entity recognition and relation extraction over structured clinical notes, with models such as BioBERT (Lee et al., 2020) and its variants demonstrating strong performance on biomedical benchmarks (Luo et al., 2022). While effective for well-defined entity types, these approaches require substantial annotated data and struggle to generalize to long, conversational inputs such as nurse dictations or patient doctor conversations (Corbeil et al., 2025a).

Subsequent work explored encoder–decoder architectures for joint extraction and generation, enabling models to produce structured outputs directly from clinical text (Kim et al., 2025; Mehta, 2025). Although these models improve flexibility, they often exhibit narrative drift and schema inconsistency (Asgari et al., 2025; Zeba et al., 2025) when applied to large, fine-grained ontologies. In nursing documentation tasks, where hundreds of fields may be sparsely expressed, such models tend to overgenerate defaults or infer unstated normal findings.

More recently, large language models (LLMs) have been applied to clinical extraction tasks through fine-tuning or prompt-based inference (Kim et al., 2025; Karim and Uzuner, 2025; Sellergren et al., 2025). Fine-tuned medical LLMs show promise in capturing domain-specific terminology, but their deployment is constrained by computational cost and limited adaptability to evolving schemas. Prompt-based approaches avoid retraining but commonly rely on single-stage, end-to-end generation, which can lead to hallucinated fields, weak evidence grounding, and poor abstention behavior in safety-critical settings. (Zeba et al., 2025) Several shared tasks in the MEDIQA series (George Michalopoulos, 2026; Corbeil et al., 2025a) have highlighted these challenges, including medical order extraction and clinical question answering. Successful systems increasingly emphasize structured outputs, explicit evidence attribution, and rule-based post-processing (Agrawal et al., 2022; Yan et al., 2026). However, most existing approaches still conflate semantic relevance determination with factual assignment, forcing models to decide what is relevant and what is true in a single step. Our work explicitly decoupling relevance judgment from value assignment. We employ LLMs as controlled reasoning components for high-recall observation extraction and continuous relevance scoring (refer Figure 1), and we note that this design aligns (Liu et al., 2025b; Adam et al., 2024) that treat LLMs as modular decision-makers rather than end-to-end generators, and it is particularly well suited to ontology-grounded nursing documentation tasks with low tolerance for inference or hallucination.

3. System Overview

Our approach is designed as a multi-agent extraction framework that converts free-text nurse dictations into structured EHR observations. Instead of relying on a single end-to-end generation, our approach orchestrates a sequence of specialized agents, each responsible for a well-defined subtask within the extraction process.

Figure 1 illustrates the overall architecture. Given a nurse dictation/transcript as input, the system proceeds through five stages: (1) observation extraction, (2) ontology candidate retrieval, (3) relevance scoring, (4) deterministic value assignment, and (5) post-processing and evaluation. All the agents use Kimi-K2 (Team et al., 2025) as LLM with different system prompts under zero shot settings.

3.1. Observation Agent

In the first stage, the LLM Agent processes the full nurse transcript to extract candidate clinical observations along with supporting evidence spans. This stage is intentionally permissive (maximizing recall) as its free form extractions tries to extrall all potential relevant observation without thinking of what ontology fields can be extracted. In the first stage, the Observation Agent (an LLM Agent) processes the full nurse transcript to extract candidate clinical observations along with their supporting evidence spans. This stage is intentionally permissive (maximizing recall), allowing free-form extraction of all potentially relevant clinical observations without considering any value specific constraints. At this point, the agent focuses solely on identifying explicit clinical facts expressed in the transcript, and not focus on value assignment, normalization, or any reasoning that would require converting or interpreting the extracted observations.

System &/ User Prompts

System prompt:
You are a clinical information extractor.

From the transcript, extract all explicit patient observations, symptoms, clinical findings, interventions, and clinician interpretations.

Rules:

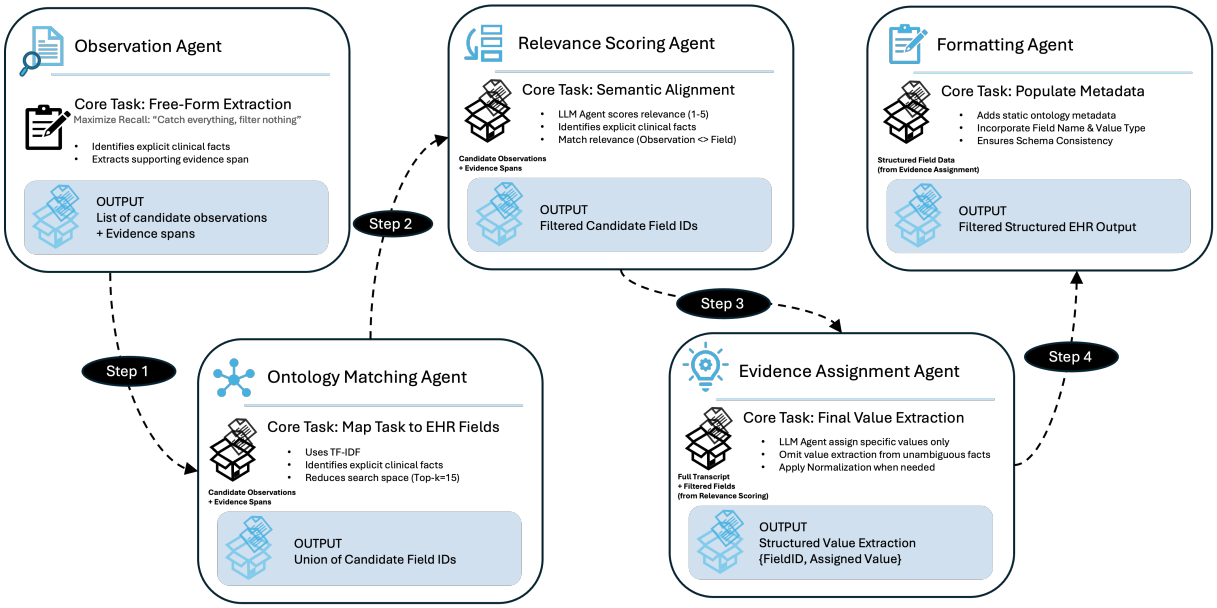


Figure 1: Overview of the LLM-driven pipeline (agent-inspired decomposition) workflow for MEDIQA-SYNUR shared task. The pipeline starts from (1) Observation Agent which extracts clinical concepts, then (2) Ontology Matching Agent reduces search space using TF-IDF retrieval, and (3) Relevance Scoring Agent filters candidate keys (score ≥ 4). Finally, outputs are generated via (4) Evidence Assignment Agent using Medprompt-inspired multi-step reasoning and finally, (5) Formatting Agent, which formats and postprocesses the outputs in the final JSON schema requested in shared task.

- Use short, precise phrases
- Stick strictly to what is stated or directly implied
- Do NOT map to any predefined schema
- Do NOT infer values that are not supported
- Output as a JSON array of atomic observations in format {"concept":...., "evidence":....}

User prompt:

**Transcript: {transcript}

Let's think step-by-step for finding Observations:

3.2. Ontology Matching Agent

The Ontology Matching Agent is responsible for mapping extracted clinical observations to a restricted set of candidate EHR fields. Given the large schema fields (192 fields), directly including all the fields in LLM calls would be inefficient and prone to hallucination/over-generation (Asgari et al., 2025). Following the work of (Zhan et al., 2021), we employ a lightweight TF-IDF–based retrieval mechanism to reduce the search space while preserving recall. For each field, we construct a textual representation consisting of the field name and when available its enumerated values. These representations form a TF-IDF document corpus using unigram and bigram features with ℓ_2 normalization. For each extracted observation, we create a query by concatenating the observation concept and its supporting evidence span. The query is projected into the same TF-IDF space, and cosine similarity is used to find the top-k fields ($k = 15$) with positive similarity scores as candidate fields. The union of retrieved field IDs across observations forms the candidate set of fields. We note that this stage significantly reduced the search space while maintaining high recall.

3.3. Relevance Scoring Agent

The Relevance Scoring Agent (an LLM Agent) evaluates the semantic alignment between extracted observations and candidate ontology fields by assigning a continuous relevance score (1 to 5; where 1 means no support and 5 means explicit transcript support) to each candidate field. We introduce this stage because LLMs, when directly asked to extract structured outputs from transcripts, often struggle to abstain and may over-generate for false/unsupported fields which eventually hurt the precision metric.

System &/ User Prompts

System prompt:
You are a clinical information extraction system.

Your task is to SCORE the relevance of candidate EHR fields.

You will be given:
1) The original clinical transcript.
2) A list of extracted clinical observations.
3) A list of candidate EHR fields with field ids.

For EACH candidate field, assign a relevance score from 1 to 5 based on how well it is supported by the transcript.

Scoring scale:
- 5 → Explicit, direct evidence in the transcript
- 4 → Strongly implied by transcript wording
- 3 → Weak, partial, or indirect evidence
- 2 → Mentioned but likely not relevant
- 1 → No meaningful support or irrelevant

Important rules:
- Prefer transcript wording over extracted observations.
- Negated findings SHOULD still receive a high score if the field is clinically relevant (negation will be handled in value extraction).
- Historical or uncertain information MAY still be scored if explicitly mentioned.
- When unsure, prefer assigning a LOWER score rather than excluding the field.
- Do NOT assign values.
- Do NOT explain your reasoning.

Output format:
- Return a valid JSON array of objects, one per candidate field:

{
"field_id": "<string>",
"score": <integer between 1 and 5>
}

Ensure every candidate field appears exactly once in the output.

Output ONLY the JSON array.

User prompt:

CLINICAL TRANSCRIPT:
{transcript}

EXTRACTED OBSERVATIONS:
{observations from step 3.2}

**CANDIDATE EHR FIELDS:
{candidate_fields}

Let's think step-by-step.

After scoring, only fields with threshold ($>= 4$) scores are retained and passed to the Evidence Assignment Agent. We note that this filtering is critical in reducing false positives, while maintaining recall for strongly supported clinical findings.

3.4. Evidence Assignment Agent

The Evidence Assignment Agent (an LLM Agent) performs the final value extraction given the full clinical transcript and the filtered candidate fields from the Relevance Scoring Agent. In this the agent is instructed to assign values only when explicit, unambiguous transcript evidence exists. Further following the recent work (Nori et al., 2023; Mehta, 2025), we enforce a mandatory multi-step decision process (spanning explicit evidence identification, controlled field value assignemtn, redundancy resolution etc). To further prevent hallucination and over-generation, the agent is explicitly instructed from inferring implicit clinical meaning or producing partially supported outputs. For categorical fields (e.g., SINGLE_SELECT or MULTI_SELECT), values must be chosen strictly from the provided enumerations. For STRING fields, short phrases are extracted from the transcript without paraphrasing.

System &/ User Prompts

System prompt:
You are a deterministic clinical information extraction system. Your sole function is to map extracted clinical observations to structured EHR fields with maximum precision and zero inference. You must operate as a fixed state machine. Follow the decision process exactly. Skipping or reordering steps is not allowed.

TASK
You will be given: 1) A list of extracted observations, each with a concept and supporting evidence. 2) A list of candidate EHR fields, each with a field id, value type, and allowed values.
Your task is to assign values to EHR fields strictly based on explicit evidence.

MANDATORY DECISION PROCESS

Step 1: Context Scan
Read all observations and candidate fields completely before assigning any value.

Step 2: Definitive Support Test
For each candidate field, check whether there exists a clear, direct, and unambiguous observation that explicitly supports that field.

EXTRACT ONLY IF:
- The observation concept or clinical evidence directly matches the field intent. - The support is explicit and literal.

IGNORE IF:
- The evidence is implied, inferred, summarized, or clinically reasoned.
- The language is tentative, speculative, explanatory, or interpretive.

- The support is partial, indirect, or distributed across multiple observations.
- The field represents a normal, negative, baseline, or default state not explicitly stated.

Step 3: Field Assignment
Assign a value ONLY if the Definitive Support Test is passed. If a field is NOT directly supported by any single observation (either by concept or by clinical evidence), you MUST OMIT it entirely. Absence of evidence is NOT evidence.

Step 4: Redundancy & Conflict Resolution
If multiple candidate fields could encode the same clinical fact, select ONLY the most specific field. Do NOT output multiple fields representing the same information.

Step 5: Abstention Bias
When evidence strength is uncertain, OMIT the field. Omission is always preferred over weak or partial assignment.

VALUE ASSIGNMENT RULES
- Use ONLY the provided observations and candidate fields.
- Do NOT answer partially supported fields.

For SINGLE_SELECT & MULTI_SELECT fields:
- Choose values STRICTLY from the provided value_enum.

For STRING fields:
- Extract a short phrase directly and verbatim from the observation evidence.
- Do NOT combine or paraphrase multiple observations.

STRICT PROHIBITIONS
- Do NOT invent fields, values, or interpretations.
- Do NOT infer clinical meaning beyond what is explicitly stated. - Do NOT infer normal, negative, improved, worsened, or default states.
- Do NOT combine multiple observations to justify a single field.
- Do NOT output field names, value types, enums, confidence, or explanations.

MANDATORY FINAL CHECK (INTERNAL)
Before returning the output, verify:
- Every field is explicitly supported by evidence.
- No inferred or default states are present.
- No redundant or overlapping fields exist.
- Output strictly follows the required JSON format.

Correct any violation before returning the final result.

OUTPUT FORMAT
Return a valid JSON array.
Each item must be exactly:
{ "id": "<field_id>", "value": <value or list of values> }

Output ONLY the JSON array.

User prompt:

You are filling structured EHR fields from extracted observations.

**Observations: {observations}

**Candidate_Fields: {Candidate_fields}

Instructions for each candidate field:

- Assign a value ONLY if supported by observations.
- SINGLE_SELECT / MULTI_SELECT: choose ONLY from value_enum.
- STRING: extract a short phrase from evidence.
- If no evidence exists, OMIT the field.
- Do NOT invent new fields or values.
- Output JSON array:
{ "id": <field_id>, "value": <value or list> }

3.5. Formatting Agent

The Formatting Agent performs post-processing to transform the Evidence Assignment Agent outputs into the final structured format required for submission. The previous stage generates only the minimal structured information (field `id` and assigned `value`), while ontology metadata such as field name and value type (e.g., STRING, SINGLE_SELECT, MULTI_SELECT) remain fixed and are not regenerated by the LLM. Since there exists a one-to-one mapping between each field ID and its associated metadata, we add them to final predictions via post-processing rather than requiring the Evidence Assignment Agent to reproduce static schema information. This design reduces generation errors and enforces schema consistency.

4. Experiments

4.1. Dataset Description

We evaluate our approach on the MEDIQA-SYNUR dataset (Corbeil et al., 2025b), which consists of free-text nurse dictations paired structured EHR observations. The task requires mapping clinical concepts from nurse dictations/transcript to a large ontology (approx. 192) distinct fields, spanning binary, categorical, and free-text value types. The test split provided by the organizers is used for final evaluation and leaderboard submission. We note that no additional external data or annotations are used.

Rank	Participant	Precision	Recall	F1
1	syhwng	0.83	0.80	0.81
2	Rezaul	0.79	0.82	0.80
3	***Smart_solutions***	**0.80**	**0.80**	**0.80**
4	avaliev	0.78	0.82	0.80
5	kyominhwang	0.79	0.81	0.80
6	singhankit16	0.73	0.82	0.77
7	krish_303	0.78	0.77	0.77
8	yudanta	0.70	0.79	0.74
9	msabanluc	0.77	0.66	0.71
10	abrygo	0.61	0.83	0.70
11	Sushvin Marimuthu	0.62	0.72	0.67
12	akabdul	0.53	0.67	0.59
13	av_dx	0.60	0.55	0.57
–	**Task Organizers**	**0.84**	**0.83**	**0.84**

Table 1: Official MEDIQA-SYNUR test set results. Rankings are shown for participating teams only.

4.2. Implementation Details

All the LLM-based agents make use of the widely cited `Kimi-K2-Instruct` (Team et al., 2025) model, which we served using vLLM(Kwon et al., 2023) on Nvidia H200 GPUs. Our approach operates in a zero-shot setting without any task-specific fine-tuning being performed. The Ontology Matching Agent performs retrieval in Stage 2 using TF-IDF–based similarity over field names and descriptions, with the top-k candidates ($k = 15$) retained per observation. We used batch processing where possible to improve throughput, while ensuring that each transcript was processed independently to preserve consistency across stages. We note that the current implementation prioritizes correctness and reliability over latency, and future work will explore optimizations and more efficient model variants to improve deployment feasibility.

4.3. Evaluation Metrics

Official MEDIQA-SYNUR (George Michalopoulos, 2026) evaluated submissions by reporting precision, recall, and F1 scores over extracted EHR fields. The Precision metric measures the proportion of predicted fields that are correct, while recall measures the proportion of gold fields that are successfully extracted, and F1 provides their harmonic mean. Given the safety-critical nature of nursing documentation, balanced precision and recall are emphasized, as both false positives and false negatives can lead to incorrect clinical records.

5. Results

In this section we report the official results of MEDIQA-SYNUR shared task in Tab 1. Our approach (Smart_solutions) ranked at third place among participating teams on the test set, with a precision of 0.80, recall of 0.80, and F1 score of 0.80.

The top participating system achieved an F1 score of 0.81, indicating that our zero-shot multi-agent pipeline performs competitively with only a marginal difference. We note that in the contrast to other approaches that favor recall at the cost of increased false positives, our approach maintains precision without sacrificing coverage, which is particularly important in clinical documentation settings. The tight clustering of scores among the top-performing systems highlights the sensitivity of the task to system design choices. We note that, despite operating in a fully zero-shot setting without any task-specific training, our pipeline achieves competitive performance. This indicates that carefully designed systems built on generalist models can serve as strong, immediately deployable baselines, particularly in settings where fine-tuning is not feasible.

Model	#Parameters	Precision	Recall	F1
GPT-OSS	120B	0.80	0.72	0.76
DeepSeek-V3.1	675B	0.77	0.66	0.71
Kimi-K2 Instruct	**1T**	**0.80**	**0.80**	**0.80**

Table 2: Zero-shot performance comparison across three widely cited open-weight LLMs under identical experimental settings.

To evaluate the impact of model choice under identical pipeline settings, we benchmark three widely cited open-weight LLMs in a fully zero-shot setting. As shown in Table 2, Kimi-K2 Instruct (1T parameters) achieves the strongest overall performance, with balanced precision and recall of 0.80. In comparison, GPT-OSS-120B (Agarwal et al., 2025) achieves competitive precision (0.80) but lower recall (0.72), resulting in an F1 score of 0.76. DeepSeek-V3.1 (Liu et al., 2024) exhibits reduced recall (0.66), leading to an F1 score of 0.71.

These results suggest that larger-scale instruction-tuned models provide improved robustness in structured clinical extraction tasks under zero-shot conditions. Notably, all models were evaluated without task-specific fine-tuning, reinforcing that performance differences arise from model capacity and instruction alignment rather than supervised adaptation.

6. Ablations

To assess the contribution of different stages in our proposed pipeline, we perform a lightweight ablation analysis by selectively removing or modifying components. The goal is to understand how different stages impacts the balance between precision and recall.

The results in Table 3 highlights the importance of decomposing the extraction process into structured stages. The single end-to-end baseline, which jointly performs observation extraction and value assignment, achieves a lower F1 score of 0.70,

Variant	Precision	Recall	F1
Single end to end call	0.73	0.68	0.70
Our Pipeline	**0.80**	**0.80**	**0.80**
w/o Relevance Scoring Agent	0.72	0.79	0.75
w/o Evidence Assignment Constraints	0.76	0.79	0.77

Table 3: Removing relevance scoring leads to increased false positives, while removing evidence constraints results in unsupported assignments.

indicating that conflating relevance determination with factual assignment leads to both missed fields and unsupported predictions.

In contrast, our full pipeline achieves balanced precision and recall (F1 = 0.80), demonstrating the effectiveness of separating these decisions. Removing the Relevance Scoring Agent leads to a significant drop in precision (0.72) while slightly increasing recall (0.82), reflecting the inclusion of unsupported fields due to the absence of explicit filtering, which supports our decision of using relevance scoring agent in final pipeline. Similarly, removing evidence-based constraints in the assignment stage results in reduced precision (0.76), as the model produces weakly grounded or partially supported outputs. Overall, these results provide empirical support for our design choice of multi agent pipelines proposed for this task.

7. Conclusion

We presented a multi-stage LLM based agentic pipeline for nurse observation extraction, as submitted to the MEDIQA-SYNUR 2026 shared task(George Michalopoulos, 2026). By decomposing the task into observation extraction, ontology-level relevance scoring, and abstention-biased value assignment, our system achieves competitive performance while minimizing hallucination and implicit inference.

Our approach placed third (refer Table 1) among participating teams on the MEDIQA-SYNUR test set, with balanced precision and recall of 0.80, despite operating in a fully zero-shot setting without task-specific training. These results demonstrate that careful system design and controlled use of LLMs can effectively address the challenges of ontology-grounded clinical documentation.

The key insight of this work is that relevance judgment and factual assignment should be treated as separate problems in clinical information extraction. Treating LLM based agents as modular decision components, rather than end-to-end generators, enables safer and more interpretable extraction pipelines. We also note that while our pipeline yields strong performance, its reliance on Kimi-K2-Instruct model ($\sim$1T parameters) adds a substantial computational overhead and may limit its practical deployment, especially in low resource settings. Future work will explore addressing such issues, and the extension of this framework to other clinical documentation tasks.

Ethics Statement and Limitations

This work focuses on automated extraction of clinical observations from nursing transcripts in a shared task setting. While the proposed pipeline aims to reduce documentation burden, it is not intended for direct clinical deployment without human oversight. All experiments were conducted on publicly available shared-task data, and no identifiable patient information was accessed outside the task guidelines. Any real-world deployment would require strict adherence to healthcare data privacy regulations and comprehensive clinical validation. Despite achieving competitive performance, several limitations remain. The pipeline relies on a large instruction-tuned model (Kimi-K2), which introduces substantial computational overhead and may limit deployment in resource-constrained environments. The multi-stage design, while improving control and interpretability, increases latency and system complexity. Additionally, the zero-shot setup may be sensitive to domain shifts and variations in clinical language. Finally, the system has not been evaluated in real-world clinical workflows, and its robustness to noisy or incomplete transcripts remains an open question.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Huda Adam, Jialing Lin, Jason Lin, Hannah Keenan, Andrew Wilson, and Marzyeh Ghassemi. 2024. Clinical information extraction with large language models: A case study on organ procurement. In *AMIA Annual Symposium Proceedings*, pages 115–123.

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*.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Fernando R Altermatt, Andres Neyem, Nicolás I Sumonte, Ignacio Villagrán, Marcelo Mendoza,

Hector J Lacassie, and Alejandro E Delfino. 2025. Evaluating gpt-4o in high-stakes medical assessments: performance and error analysis on a chilean anesthesiology exam. *BMC Medical Education*, 25(1):1499.

Ehsaneddin Asgari, Natalia Montaña-Brown, Marie Dubois, Sally Khalil, James Balloch, Jonathan A. Yeung, and Daniela Pimenta. 2025. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):274.

Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. 2024. Med42–evaluating fine-tuning strategies for medical llms: full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*.

Jean-Philippe Corbeil, Asma Ben Abacha, Jérôme Tremblay, Phillip Swazinna, Akila Jeeson Daniel, Miguel Del-Agua, and François Beaulieu. 2025a. Overview of the mediqa-oe 2025 shared task on medical order extraction from doctor-patient consultations. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 11–16.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025b. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Suzhou (China). Association for Computational Linguistics.

Adriano Friganović, Polona Selič, Biljana Ilić, and Biserka Sedić. 2019. Stress and burnout syndrome and their associations with coping and job satisfaction in critical care nurses: a literature review. *Psychiatria Danubina*, 31(Suppl 1):21–31.

Cari Bader Nate Bodenstab Asma Ben Abacha George Michalopoulos, Jean-Philippe Corbeil. 2026. Overview of the mediqa-synur 2026 shared task on observation extraction from nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiacheng Hu, Runyuan Bao, Yang Lin, Hanchao Zhang, and Yanlin Xiang. 2024. Accurate medical named entity recognition through specialized nlp models. In *2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC)*, pages 578–582. IEEE.

Thelma C Hurd, Fay Cobb Payton, Darryl B Hood, et al. 2024. Targeting machine learning and artificial intelligence algorithms in health care to reduce bias and improve population health. *The Milbank Quarterly*, 102(3):577.

AHM Karim and Ozlem Uzuner. 2025. Assessing large language models for structured medical order extraction. *arXiv preprint arXiv:2510.10475*.

Min Soo Kim, Paul Chung, Nima Aghaeepour, and Namkug Kim. 2025. Information extraction from clinical texts with generative pre-trained transformer models. *International Journal of Medical Sciences*, 22(5):1015–1028.

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*.

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.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Hui Liu, Ziyi Chen, Peilin Li, Yuan-Zhi Liu, Xiangtao Liu, Ronald X Xu, and Mingzhai Sun. 2025a. Resource-efficient instruction tuning of large language models for biomedical named entity recognition. *Journal of Biomedical Informatics*, page 104896.

Longchao Liu, Long Lian, Yiyan Hao, Aidan Pace, Elaine Kim, Nour Homsi, Yash Pershad, Liheng Lai, Thomas Gracie, Ashwin Kishtagari, et al. 2025b. Human level information extraction from

clinical reports with finetuned language models. *Scientific Reports*, 15(1):45239.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Zoltan P Majdik, S Scott Graham, Jade C Shiva Edward, Sabrina N Rodriguez, Martha S Karnes, Jared T Jensen, Joshua B Barbour, and Justin F Rousseau. 2024. Sample size considerations for fine-tuning large language models for named entity recognition tasks: methodological study. *Jmir ai*, 3:e52095.

Matthew D. McHugh, Ann Kutney-Lee, Jeannie P. Cimiotti, Douglas M. Sloane, and Linda H. Aiken. 2011. Nurses' widespread job dissatisfaction, burnout, and frustration with health benefits signal problems for patient care. *Health Affairs*, 30(2):202–210.

Parth Mehta. 2025. Pnlp at mediqa-oe 2025: A zero-shot prompting strategy with gemini for medical order extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 75–83.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

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 informatics*, 32(1):e101139.

Cecilia Padilla Fortunatti and Yaritza K. Palmeiro-Silva. 2017. Effort-reward imbalance and burnout among icu nursing staff: A cross-sectional study. *Nursing Research*, 66(5):410–416.

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*.

Soorya Ram Shimgekar, Shayan Vassef, Abhay Goyal, Navin Kumar, and Koustuv Saha. 2025. Agentic ai framework for end-to-end medical data inference. *arXiv preprint arXiv:2507.18115*.

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.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Robin Willard-Grace, Margaret Knox, Bruce Huang, Heather Hammer, Chris Kivlahan, and Kevin Grumbach. 2019. Burnout and health care workforce turnover. *Annals of Family Medicine*, 17(1):36–41.

Qianqi Yan, Huy Nguyen, Sumana Srivatsa, Hari Bandi, Xin Eric Wang, and Krishnaram Kenthapadi. 2026. Cite-while-you-generate: Training-free evidence attribution for multimodal clinical summarization. *arXiv preprint arXiv:2601.16397*.

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*.

Musarrat Zeba, Abdullah Al Mamun, Kishoar Jahan Tithee, Debopom Sutradhar, Mohaimenul Azam Khan Raiaan, Saddam Mukta, Reem E Mohamed, Md Rafiqul Islam, Yakub Sebastian, Mukhtar Hussain, et al. 2025. Mitigating hallucinations in healthcare llms with granular fact-checking and domain-specific adaptation. *arXiv preprint arXiv:2512.16189*.

Xiaoxuan Zhan, Michael Humbert-Droz, Partha Mukherjee, and Olivier Gevaert. 2021. Structuring clinical text with AI: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. *Patterns*, 2(7):100289.

Zhiyu Zhang and Arbee LP Chen. 2022. Biomedical named entity recognition with the combined feature attention and fully-shared multi-task learning. *BMC bioinformatics*, 23(1):458.

GS-BrainText: A Multi-Site Brain Imaging Report Dataset from Generation Scotland for Clinical Natural Language Processing Development and Validation

Beatrice Alex[1,2,3,9], **Claire Grover**[4], **Arlene Casey**[1,5], **Richard Tobin**[4], **Heather Whalley**[6,7], **William Whiteley**[6,8]

[1]Advanced Care Research Centre, University of Edinburgh, Edinburgh, UK
[2]School of Literatures, Languages and Cultures, University of Edinburgh, UK
[3]Edinburgh Futures Institute, University of Edinburgh, UK
[4]School of Informatics, University of Edinburgh, UK
[5]Usher Institute, University of Edinburgh, UK
[6]Institute for Neuroscience and Cardiovascular Research, University of Edinburgh, UK
[7]Generation Scotland, Institute of Genetics and Cancer, University of Edinburgh, UK
[8]NHS Lothian, Edinburgh, UK
[9]School of Mathematical and Computer Sciences, Heriot-Watt University, UK
Corresponding author: b.alex@ed.ac.uk

Abstract

We present GS-BrainText, a curated dataset of 8,511 brain radiology reports from the Generation Scotland cohort, of which 2,431 are annotated for 24 brain disease phenotypes. This multi-site dataset spans five Scottish NHS health boards and includes broad age representation (mean age 58, median age 53), making it uniquely valuable for developing and evaluating generalisable clinical natural language processing (NLP) algorithms and tools. Expert annotations were performed by a multidisciplinary clinical team using an annotation schema, with 10–100% double annotation per NHS health board and rigorous quality assurance. Benchmark evaluation using EdIE-R, an existing rule-based NLP system developed in conjunction with the annotation schema, revealed some performance variation across health boards (F1: 86.13–98.13), phenotypes (F1: 22.22–100) and age groups (F1: 87.01–98.13), highlighting critical challenges in generalisation of NLP tools. The GS-BrainText dataset addresses a significant gap in available UK clinical text resources and provides a valuable resource for the study of linguistic variation, diagnostic uncertainty expression and the impact of data characteristics on NLP system performance.

Keywords: clinical natural language processing, language resources, brain imaging reports, radiology dataset, gold standard, benchmarking

1. Introduction

Clinical natural language processing (NLP) can unlock valuable information from unstructured electronic health records, supporting research, clinical decision-making and healthcare quality improvement. However, the translation of NLP tools from research to clinical practice remains challenging, primarily due to limited access to diverse, representative datasets for development and validation.

Current clinical NLP research faces several critical limitations. First, most publicly available clinical text datasets originate from US healthcare systems and may not generalise to other countries, which differ in terminology, documentation practices and healthcare processes. Second, existing datasets often represent single institutions or narrow patient populations, limiting their utility for developing robust, generalisable NLP systems. Third, the highly sensitive nature of clinical data creates significant barriers to data sharing, hindering external validation of NLP tools.

We present GS-BrainText, a curated dataset of brain radiology reports from Generation Scotland, a broad population-based cohort spanning multiple Scottish NHS health boards. This dataset addresses key gaps in available clinical NLP resources by providing: (1) multi-site representation across various Scottish healthcare settings, (2) broad demographic coverage including age ranges typically under-represented in clinical datasets, (3) expert annotations for clinically important cerebrovascular phenotypes and (4) evidence of NLP system performance variation across phenotypes, sites and age ranges.

2. Background

2.1. Generation Scotland

Generation Scotland is a population-based health study of approximately 24,000 participants from the Scottish Family Health Study and the 21st Century Genetic Health study (Smith et al., 2006; Milbourn et al., 2024). Established to investigate genetic and environmental contributions to common diseases,

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 92–102
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

participants consented to data linkage and sharing of their health records, including radiology reports. The cohort spans multiple NHS health boards serving distinct geographic regions in Scotland with varying demographic characteristics. Unlike single-site or disease-specific cohorts, Generation Scotland represents a population-based sample that includes individuals (recruited between 2006 and 2010) across the full adult age spectrum.[1]

As part of routine clinical care, participants in Generation Scotland who underwent brain imaging (CT or MRI) have their radiology reports stored in NHS digital systems. These reports, authored by consultant radiologists, document findings observed in brain scans. Reports follow standard radiology reporting conventions, with sections for clinical history (provided by the referring clinician), observations and opinion. The length of the report varies from a few words to a maximum of 602 words (mean: 82 words) depending on the indication of the scan, the practice of the radiologist, and the complexity of the findings. Routine clinical reports are primarily written for clinical communication. The GS-BrainText dataset includes reports spanning 1994–2021, providing longitudinal imaging data for the Generation Scotland cohort.

A number of research study reports are also included in GS-BrainText but marked in the metadata to be excludable if necessary. Most of these study reports (784) are in the unannotated subset, with 14 in the annotated subset.

2.2. Brain Disease Phenotypes

Phenotypes in clinical NLP refer to observable characteristics or traits that can be identified from medical records. In the context of brain radiology reports, phenotypes represent specific clinical findings or diagnostic categories that radiologists observe and document when interpreting brain imaging. For example, they include conditions such as ischaemic stroke (with specifications for location and time of occurrence), small vessel disease, atrophy and various types of tumours.

Automated phenotype extraction from unstructured radiology text enables researchers to identify patient cohorts with specific characteristics, conduct large-scale epidemiological studies and investigate disease patterns that would be infeasible to detect through manual review alone. The accurate identification and classification of these phenotypes from free-text reports is essential for translating imaging findings into structured, searchable data that can support both clinical care and research.

We focus on 24 brain disease phenotypes with substantial clinical and research importance:

Stroke-Related Labels (11)

1. Ischaemic stroke, deep, recent
2. Ischaemic stroke, deep, old
3. Ischaemic stroke, cortical, recent
4. Ischaemic stroke, cortical, old
5. Ischaemic stroke, underspecified
6. Haemorrhagic stroke, deep, recent
7. Haemorrhagic stroke, deep, old
8. Haemorrhagic stroke, lobar, recent
9. Haemorrhagic stroke, lobar, old
10. Haemorrhagic stroke, underspecified
11. Stroke, underspecified

Tumour-Related Labels (4)

12. Tumour, meningioma
13. Tumour, metastasis
14. Tumour, glioma
15. Tumour, other

Other Neurological Findings (9)

16. Small vessel disease
17. Atrophy
18. Subdural haematoma
19. Subarachnoid haemorrhage, aneurysmal
20. Subarachnoid haemorrhage, other
21. Microbleed, deep
22. Microbleed, lobar
23. Microbleed, underspecified
24. Haemorrhagic transformation

These phenotypes are annotated as document-level labels (see Section 4.1). They were first devised as part of earlier work on text mining brain imaging reports by Alex et al. (2019) using study specific data from one Scottish NHS health board (the Edinburgh Stroke Study, (Jackson et al., 2008)), which was motivated by the fact that existing medical ontological entries were not sufficiently fine-grained and phenotyping using image analysis was not sufficiently accurate yet. The document-level labels combine mark-up in the text itself for observations (like ischaemic stroke or tumour) as well as modifiers for location in the brain (deep/cortical/lobar) and time of occurrence (old/recent), creating a comprehensive taxonomy for stroke, tumour and other neurological findings from brain imaging reports. Automated extraction of these phenotypes enables large-scale research studies that would be infeasible through manual review alone (e.g., Camilleri et al., 2025; Iveson et al., 2025).

[1] Generation Scotland now also includes an additional 17,000 participants from across the whole of Scotland (NextGenScot), including ones who are younger and from more health boards. Their data is not included in GS-BrainText.

2.3. Existing Datasets and Tools

Despite the critical need for annotated clinical text datasets to develop and validate NLP systems, publicly available annotated brain imaging report datasets remain scarce. While some large-scale annotated radiology datasets exist, such as MIMIC-CXR (Johnson et al., 2019), containing over 377,000 chest X-ray reports, and IU-XRay, the Indiana University Chest X-ray collection (Demner-Fushman et al., 2016), these focus on chest imaging rather than neuroimaging and originate from US healthcare systems. Linguistic conventions, terminology and documentation practices in US radiology reports differ from UK NHS settings, limiting their utility for UK clinical deployment.

Recently, two major brain imaging datasets have been introduced. Camilleri et al. (2025) describe a large-scale national cohort from NHS Scotland (approx. 1.3M brain imaging reports/images linked to health systems data) accessible through the Brain Health Data Pilot with reports being labelled with EdIE-R system output.[2] The Brain Imaging and Neurophysiology Database (BIND), approx. 85K radiology reports from over 1.8M neuroimaging scans across three US academic medical centres, used biomedical large language models (LLM) to extract structured clinical metadata from unstructured reports (Maschke et al., 2025). However, the authors explicitly note that due to LLM limitations and the absence of human-curated annotations, they "cannot guarantee complete accuracy of extracted information", positioning the metadata as a navigation tool rather than clinical ground truth. Both datasets provide scale but lack expert manual annotations of clinical phenotypes, limiting their utility for NLP systems requiring gold standard labels.

For expertly annotated brain imaging datasets, several smaller resources exist but with restricted access and limited scope. The Edinburgh Stroke Study dataset used to develop EdIE-R comprises 1,168 manually annotated reports (Alex et al., 2019), and datasets used to develop ALARM+, ESPRESSO, and Sem-EHR remain internal to their respective institutions (Schrempf et al., 2020; Fu et al., 2019; Rannikmäe et al., 2021). They provide high-quality expert annotations for specific phenotypes but are limited in scale, geographic diversity and availability to the broader research community.

This scarcity of shareable, expertly annotated, UK-based brain radiology datasets creates significant barriers to advancing clinical NLP research for neuroimaging applications. GS-BrainText addresses this gap by providing a multi-site, population-based resource with expert clinical gold annotations for UK NHS health boards.

[2] https://healthdatagateway.org/en/collection/158

3. Data Collection and Curation

3.1. Report Acquisition

We obtained access to brain imaging reports for Generation Scotland participants who underwent brain CT or MRI scans. Reports were retrieved through established data linkage protocols with NHS Scotland health boards: Fife, Lothian, Greater Glasgow and Clyde (GGC), Grampian and Tayside.

3.2. Pre-processing Pipeline

Raw reports and associated metadata were supplied in CSV format from each health board. We developed a standardised pre-processing pipeline comprising:

Format conversion and structure identification: Reports were converted from spreadsheets to XML format using components of the EdIE-R text processing pipeline (Alex et al., 2019). During conversion, we automatically identified structural sections within reports, including clinical history and report body. Clinical history sections, which typically contain referring clinician notes rather than radiologist findings, were not annotated.

Imaging modality labelling: Reports were labelled by their modality (CT, MRI, etc.) in the XML.

Following this pre-processing, the final dataset, called the GS-BrainText data, comprises 8,511 scan reports (CT: 4362, MRI: 966, Other: 183) distributed across the five health boards: Tayside (3,747), GGC (3,309), Grampian (1,076), Lothian (269) and Fife (110). The distribution reflects the volume of reports originally provided by each board, with larger contributions from Tayside and GGC. From these available reports, we annotated 2,431 reports (CT: 1,487, MRI: 944), including all but 2 reports from Fife (108), approximately half from Lothian (143) and Grampian (521), a third from GGC (1,009) and a fifth from Tayside (650).

4. Annotation Schema and Process

4.1. Annotation Schema Design

We employed a comprehensive annotation schema originally developed together with a neurologist and a radiologist for the development of the EdIE-R clinical NLP pipeline (Alex et al., 2019; Wheater et al., 2019). It comprises of text-level annotations of entities and modifiers, attributes of entities, relations as well as document-level labels:

Text-level annotations include (1) named entities representing clinical findings (e.g., “ischaemic stroke”, “small vessel disease”, “atrophy”); (2) Location modifiers specifying anatomical regions (i.e., “cortical”, “lobar” or “deep”) and (3) temporal modifiers indicating timing (i.e., “old” or “recent”).

Attributes are an additional annotation on entities reserved for negation (i.e., to signal that a type of stroke or tumour is not visible on the image.
Relations include links between modifiers and named entities (e.g., a link between "old" and "ischaemic stroke" to signal that this type of stroke was not recent or between "deep" and "haemorrhagic stroke" to signal the location of the haemorrhage. Entities can be related to multiple modifiers.
Document-level labels classify each report for all phenotypes (see Section 2.2) in the context of all other annotations.

The schema was designed to capture not only phenotype presence and absence but also critical contextual information that affects clinical interpretation. The annotation was carried out using the BRAT annotation tool (Stenetorp et al., 2012) and a synthetic annotated example is shown in Figure 1.

To ensure consistency and efficiency, annotation was performed as systematic correction of EdIE-R output rather than annotation from scratch. As a result, residual anchoring bias cannot be excluded; in particular, false negatives not surfaced by EdIE-R may be under-represented in the gold standard. Nevertheless, pre-annotation with EdIE-R substantially reduced annotation time, a pragmatic trade-off that is commonly adopted in clinical NLP data annotation work and one taken in light of EdIE-R's previously demonstrated performance on Scottish radiology report data (see Section 6.1.

4.2. Annotation Team and Training

Annotations were performed by a multidisciplinary team consisting of a consultant neurologist, a clinical fellow, a junior doctor, a clinical laboratory scientist and a medical student. All annotators underwent comprehensive training on 115 files from the Tayside and Fife subsets through an iterative process.

4.3. Quality Assurance

We implemented rigorous quality assurance to ensure annotation consistency:

- Regular consensus meetings were held where annotators could discuss challenging cases, resolve queries and provide feedback to maintain consistency across all five annotators.
- Around 10% of data was double-annotated across NHS GGC, Grampian and Tayside.
- 100% of the NHS Lothian and Fife data was annotated by two or more annotators as we had less data available from these health boards.
- Inter-annotator agreement (IAA) was computed amongst the four less experienced annotators, i.e. not including the consultant neurologist (see Table 1).
- The final annotated dataset comprises single-annotated files and, for double-annotated files, annotations selected based on annotator expertise and consistency.

Table 1 presents IAA statistics for document-level labels (phenotypes). F1-scores range from 83.74 to 95.65 across phenotypes and health boards, indicating reasonably high or very high agreement.

IAA was assessed at both the text level (named entities, negation attributes, modifiers and relations) and the document level (phenotype labels) in GS-BrainText. However, given this paper's focus on dataset utility for phenotype extraction, we only report document-level agreement statistics. Comprehensive text-level annotation statistics, albeit for other datasets, can be found in Alex et al. (2019) and Wheater et al. (2019).

5. Annotated Dataset Characteristics

5.1. Size and Composition

The annotated GS-BrainText comprises 2,431 radiology reports across five NHS health boards. Table 2 shows the distribution of reports and phenotype prevalence by health board and demonstrates substantial variation in frequencies.[3] Small vessel disease (n=545) and atrophy (n=441) were the most common findings, together accounting for over 53% of all labels. Ischaemic stroke phenotypes show reasonable representation (n=381 total) and were more common than haemorrhagic stroke (n=84 total), consistent with known stroke epidemiology. Several clinically important phenotypes were rare in the dataset: microbleeds (n=12), haemorrhagic transformation (n=10) and gliomas (n=8). This imbalanced distribution reflects both the population-based nature of the cohort and the relative epidemiological frequencies of these conditions, but poses challenges for evaluating NLP system performance on rare phenotypes.

The distribution varies also substantially by health board, with GGC contributing the largest absolute numbers across most phenotypes due to its larger sample size. Rarer phenotypes like microbleeds appear almost exclusively in data from NHS GGC and Grampian.

[3]We do not provide a fixed data split, as the annotated subset was used solely for evaluation and several phenotypes are rare in the data. We therefore recommend future work adopt k-fold cross-validation to determine performance of the entire dataset.

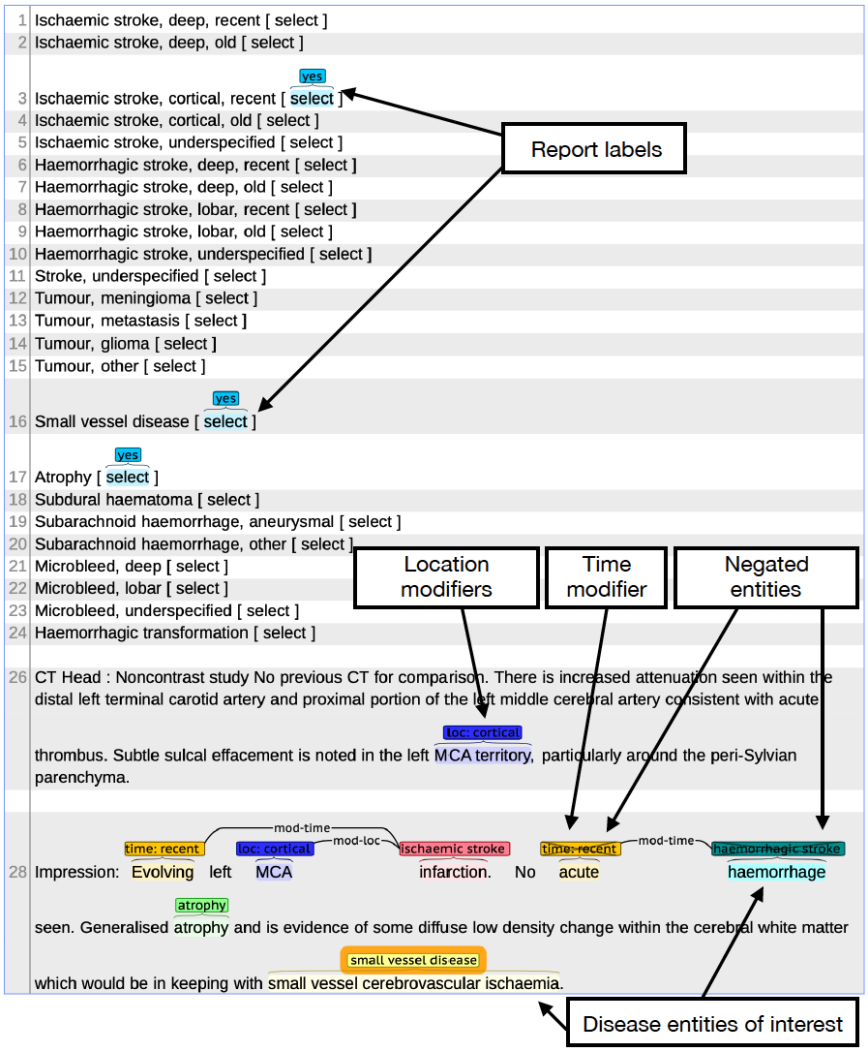


Figure 1: A synthetic brain imaging report annotated with text- and document-level annotations.

Annotator pairs	Fife	GGC	Grampian	Lothian	Tayside
A1-A2	88.15 (108)	-	-	-	-
A1-A3	87.50 (108)	-	-	-	-
A1-A4	89.90 (108)	-	-	-	-
A2-A3	83.58 (108)	91.18 (53)	-	86.51 (143)	90.91 (35)
A2-A4	90.82 (108)	-	95.65 (26)	-	-
A3-A4	86.17 (108)	86.49 (53)	91.67 (24)	-	87.50 (30)

Table 1: Pairwise inter-annotator agreement (IAA) reported in F1 (and number of reports in parentheses) for phenotype annotations in GS-BrainText across five Scottish NHS health boards (GGC = Greater Glasgow and Clyde). All annotators annotated all the reports in Fife and two annotated all the reports in Lothian. IAA for other boards was computed on smaller subsets with two pairwise comparisons each.

5.2. Demographic Distribution

The annotated dataset contains 2,431 scan reports, of which 1,433 are of female and 988 of male patients, matching the overall Generation Scotland cohort gender distribution: 59% female and 41% male (Smith et al., 2012).

Of the annotated reports, 644 are from patients under 50, 1,063 from those aged 50–69, and 704 from those aged over 70.

5.3. Linguistic Characteristics

Casey et al. (2023) previously identified important linguistic characteristics in a subset of the GS-BrainText dataset related to diagnostic uncertainty

Phenotype	Fife	GGC	Grampian	Lothian	Tayside	Total
Ischaemic_stroke_deep_recent	0	9	9	3	0	21
Ischaemic_stroke_deep_old	6	78	55	15	25	179
Ischaemic_stroke_cortical_recent	6	22	9	5	5	47
Ischaemic_stroke_cortical_old	6	45	24	5	8	88
Ischaemic_stroke_underspecified	4	17	11	0	14	46
Haemorrhagic_stroke_deep_recent	0	0	1	1	0	2
Haemorrhagic_stroke_deep_old	1	3	1	0	2	7
Haemorrhagic_stroke_lobar_recent	1	4	1	1	3	10
Haemorrhagic_stroke_lobar_old	1	8	1	0	1	11
Haemorrhagic_stroke_underspecified	6	28	11	3	6	54
Stroke_underspecified	2	3	2	0	1	8
Tumour_meningioma	7	18	18	2	3	48
Tumour_metastasis	0	22	6	4	2	34
Tumour_glioma	0	4	3	1	0	8
Tumour_other	3	24	10	5	9	51
Small_vessel_disease	26	250	178	42	49	545
Atrophy	32	215	114	21	59	441
Subdural_haematoma	0	36	25	7	0	68
Subarachnoid_haemorrhage_aneurysmal	0	2	4	1	0	7
Subarachnoid_haemorrhage_other	1	9	15	5	3	33
Microbleed_deep	0	1	1	0	0	2
Microbleed_lobar	0	3	4	0	0	7
Microbleed_underspecified	0	3	0	0	0	3
Haemorrhagic_transformation	5	4	1	0	0	10

Table 2: Distribution of phenotype prevalence across five Scottish NHS health boards in the annotated subset of GS-BrainText (GGC = Greater Glasgow and Clyde).

in radiological reporting. Their analysis showed that scan reports of younger patients frequently employ hedging language, particularly for cerebrovascular findings that are less expected in this age group. Phrases such as "this may possibly be an ischaemic infarct" appear more frequently in reports of patients under 50 years compared to those over 70. Uncertainty expressions pose challenges both for manual annotation and NLP systems.

6. Benchmark Evaluation

In prior work, Casey et al. (2023) compared the performance of four NLP systems on CT scans in GS-BrainText: EdIE-R (Alex et al., 2019; Wheater et al., 2019), ALARM+ (Schrempf et al., 2020, 2021), ESPRESSO (Fu et al., 2019, 2020) and SemEHR (Rannikmäe et al., 2021). These systems represent diverse approaches (rule-based and neural) and training or development data sources. Evaluation was restricted to a subset of three overlapping phenotypes (ischaemic stroke, small vessel disease and atrophy).

In the following section, we will lay out the performance for EdIE-R across all 24 phenotypes on the full manually annotated GS-BrainText gold data (CT and MRI). The aim is to demonstrate the usefulness of the GS-BrainText dataset providing a baseline for the entire phenotype annotation scheme for future research using machine learning or large language model approaches.

6.1. System Description

EdIE-R (short for Edinburgh Information Extraction for Radiology) is a rule-based system developed on the Edinburgh Stroke Study (ESS) data from NHS Lothian (Alex et al., 2019) and subsequently refined and evaluated on NHS Tayside data (Wheater et al., 2019).[4] After some initial preprocessing and linguistic analysis, it performs named entity recognition (NER), negation detection and relation extraction before conducting document-level classification. Its NER and negation detection components were robustly evaluated on ESS and Tayside data (Grivas et al., 2020; Sykes et al., 2021).

6.2. Evaluation

Table 3 presents EdIE-R's precision, recall and F1 scores across all 24 phenotypes on the manually annotated GS-BrainText data.[5] EdIE-R demonstrates high overall performance on this dataset,

[4] EdIE-R is available at https://www.ltg.ed.ac.uk/software/edie-r.

[5] In clinical research, precision is also called positive predictive value and recall is known as sensitivity.

achieving a micro-average F1 score of 88.82. However, micro-averaging is dominated by high-frequency phenotypes; we therefore also report macro-averaged precision, recall and F1 of 74.22, 84.78 and 77.47 respectively, which weights all phenotypes equally and reflects performance across the full range of phenotype frequencies.

System performance varies considerably by phenotype type and frequency, achieving best results on high-frequency phenotypes: small vessel disease (F1=95.6), atrophy (F1=95.69) and subdural haematoma (F1=91.55). For ischaemic stroke subtypes, performance ranges from 74.42 (deep, recent) to 91.3 (cortical, recent), with the underspecified category performing notably lower (F1=60) due to reduced precision (51.56). Microbleeds (deep and lobar) achieve perfect scores (F1=100), though their extreme rarity (n=2 and n=7) limits the reliability of these estimates. The poorest performance is obtained for subarachnoid haemorrhage, aneurysmal (F1=22.22), reflecting both the low frequency of this phenotype (n=7) and its very low recall (14.29).

Individual results for low-support phenotypes (n<30 in the Total column of Table 2) should be interpreted with caution, as scores may not be reliable estimates of generalised system performance.

6.3. Performance by Health Board

EdIE-R performance across health boards (see Table 4) reveals some variation, with F1 scores ranging from 86.13 (Tayside) to 98.13 (Fife). The system achieves highest performance on Lothian (96.36) and Fife (98.13). Notably, Lothian data (though unlikely the specific reports in GS-BrainText) were used in EdIE-R's original development on the Edinburgh Stroke Study (Alex et al., 2019), while data from Fife was not used when devising its rules. Performance is lower on other health boards: GGC (87.01), Grampian (89.15) and Tayside (86.13). The reduced performance on Tayside is particularly noteworthy, as similar NHS Tayside data was used for system refinement (Wheater et al., 2019). The variation in F1 scores demonstrates the impact of local reporting conventions and data heterogeneity on rule-based system performance, even within a single national healthcare system.

6.4. Performance by Age Group

Performance increased with age (ranging between F1=77.60 for <50 and F1=91.7 for 70+), corresponding to phenotype coverage of 16/24 for people under 50, 24/24 for people aged 50-69, and 23/24 for those aged over 70 (see Table 5).

7. Discussion and Conclusion

We present BS-BrainText, a dataset of 8,511 brain radiology reports from Generation Scotland, of which 2,431 are annotated for 24 cerebrovascular phenotypes across five Scottish NHS health boards. Benchmark evaluation of the EdIE-R NLP system reveals strong overall performance (F1=88.82) with substantial variation across phenotypes (22.22–100), health boards (86.13–98.13) and age (77.6–91.7), highlighting challenges for clinical NLP deployment. Performance exceeds 95 in F1 for common phenotypes, but drops for rare ones and differs across NHS sites, even within a single national healthcare system.

The observed variation highlights critical factors affecting clinical NLP robustness. Local reporting conventions substantially impact performance: EdIE-R achieves 96.36 on Lothian data similar to what was used in its development, yet performs at 86–89 on data from some other boards. Interestingly, performance on Fife (98.13) exceeds that on Tayside (86.13) despite similar Tayside data having been used for system refinement. Its performance variation of up to 12 points across sites underscores the challenges of developing truly generalisable clinical NLP systems.

These findings carry important implications for clinical deployment. "Out-of-the-box" application of NLP systems to new healthcare settings, even within the same country, carries some risk of performance degradation. Organisations adopting NLP tools should therefore conduct local validation before deployment and monitor performance across demographic and clinical subgroups. For researchers, our results underscore the value of multi-site datasets for developing robust systems, as training on single-institution data does not prepare systems for linguistic and clinical variation encountered across sites.

GS-BrainText, containing expert annotations across 24 clinically important phenotypes, provides a valuable resource for advancing robust and generalisable clinical NLP, enabling researchers to develop systems that can handle the linguistic variation, demographic diversity and multi-site heterogeneity encountered in real-world healthcare settings. Through controlled access, this dataset can support the development of next-generation clinical NLP tools that maintain reliable performance across the diverse contexts of clinical practice.

With EdIE-R, we contribute a strong baseline (micro-average F1=88.82), demonstrating that well-engineered rule-based approaches can achieve competitive performance on clinical phenotyping tasks while providing interpretable predictions for benchmarking new approaches.

Phenotype	Precision	Recall	F1
Ischaemic_stroke_deep_recent	72.73	76.19	74.42
Ischaemic_stroke_deep_old	84.58	94.97	89.47
Ischaemic_stroke_cortical_recent	93.33	89.36	91.30
Ischaemic_stroke_cortical_old	86.02	90.91	88.40
Ischaemic_stroke_underspecified	51.56	71.74	60.00
Haemorrhagic_stroke_deep_recent	66.67	100.00	80.00
Haemorrhagic_stroke_deep_old	75.00	42.86	54.55
Haemorrhagic_stroke_lobar_recent	77.78	70.00	73.68
Haemorrhagic_stroke_lobar_old	72.73	72.73	72.73
Haemorrhagic_stroke_underspecified	57.14	96.30	71.72
Stroke_underspecified	36.84	87.50	51.85
Tumour_meningioma	82.76	100.00	90.57
Tumour_metastasis	64.00	94.12	76.19
Tumour_glioma	66.67	75.00	70.59
Tumour_other	54.05	78.43	64.00
Small_vessel_disease	93.66	97.61	95.60
Atrophy	98.33	93.20	95.69
Subdural_haematoma	87.84	95.59	91.55
Subarachnoid_haemorrhage_aneurysmal	50.00	14.29	22.22
Subarachnoid_haemorrhage_other	63.27	93.94	75.61
Microbleed_deep	100.00	100.00	100.00
Microbleed_lobar	100.00	100.00	100.00
Microbleed_underspecified	75.00	100.00	85.71
Haemorrhagic_transformation	71.43	100.00	83.33
Micro-average	85.03	92.95	88.82
Macro-average	74.22	84.78	77.47

Table 3: Precision, recall and F1 scores for the EdIE-R NLP system on GS-BrainText for individual phenotypes as well as micro- and macro-averaged performance scores.

Health Board	Precision	Recall	F1
Fife	98.13	98.13	98.13
GGC	82.18	92.45	87.01
Grampian	87.12	91.27	89.15
Lothian	94.44	98.35	96.36
Tayside	80.09	93.16	86.13

Table 4: EdIE-R performance in precision, recall and micro-average F1-scores for GS-BrainText per NHS health board.

Age Group	Precision	Recall	F1
<50	69.78	87.39	77.60
50-69	81.24	92.76	86.62
70+	89.78	93.70	91.70

Table 5: EdIE-R performance in precision, recall and micro-average F1-scores by age group.

8. Limitations and Future Work

While GS-BrainText addresses important gaps in available clinical NLP resources, several limitations and opportunities for future work should be noted:

Geographic scope: The dataset is limited to Scottish NHS health boards and may not represent clinical reporting practices in other UK regions or international healthcare systems. Variations in terminology, reporting conventions and clinical workflows may limit generalisability beyond Scotland, though the multi-site nature within Scotland provides some indication of cross-institutional robustness.

Imaging modality: The dataset includes both CT (1,487 annotated, 4,362 total) and MRI (944 annotated, 3,966 total) reports. The predominance of CT in the annotated subset reflects its prevalence in acute clinical settings, though the substantial MRI representation (39% of annotated reports) enables investigation of cross-modality language variation. Terminology and reporting conventions may differ between CT and MRI, presenting both challenges and opportunities for developing robust cross-modality systems.

Phenotype coverage: Although we annotated 24 clinically important phenotypes, many other neurological findings documented in radiology reports are not captured by our annotation schema. The schema focuses on cerebrovascular pathology and common structural abnormalities, excluding conditions such as infectious processes and traumatic injuries beyond subdural haematoma.

Class imbalance: Substantial imbalance exists in phenotype frequencies, with some conditions (small vessel disease, atrophy) well-represented while others (microbleeds, certain tumour subtypes) being extremely rare. This imbalance, while reflecting real-world epidemiology, limits the ability to robustly evaluate NLP system performance on rare but clinically important phenotypes.

Annotation reporting: Document-level labels represent the primary annotation layer reported on in this paper, with text-level annotations (named entities, relations, temporal and location modifiers and negation) not systematically reported here. While phenotype annotation aligns with typical use cases for cohort identification and epidemiological research, researchers interested in developing entity recognition or relation extraction components can acquire additional detailed annotation statistics from the GS-BrainText data itself.

Temporal factors: Reports span over a large time frame (1994–2021; most common scan date in 2016) during which reporting practices, imaging technology and clinical guidelines have evolved. We do not systematically analyse temporal trends in language use or diagnostic patterns, though such changes may contribute to observed performance variation across subsets of the data and can be explored in future research.

Population characteristics: Generation Scotland is a volunteer cohort that may not fully represent the general Scottish population. Participants willing to engage in long-term clinical research may differ from non-participants in their readiness to seek medical help, their socioeconomic status or their baseline health. The population-based sampling reduces some selection biases present in disease-specific cohorts, but perfect representativeness cannot be assumed.

NLP system coverage: We evaluated only EdIE-R, a rule-based system. Modern contextual models may improve on some challenges, but evidence suggests site-specific variation persists across NLP architectures (Casey et al., 2023), indicating that architectural advances alone cannot eliminate the need for local validation.

Despite these limitations, GS-BrainText represents a valuable available resources for clinical NLP research on brain imaging reports, providing multi-site, expertly annotated data with comprehensive baseline performance reporting and interesting opportunities for future research.

9. Data Access

GS-BrainText is available via Generation Scotland's controlled access process: https://www.ed.ac.uk/generation-scotland/for-researchers/access.

10. Ethics Statement

This research was conducted under ethical approvals for the Generation Scotland cohort. GS has ethical approval for the SFHS study (05/S1401/89) and 21CGH study (06/S1401/27) and both studies are now part of a Research Tissue Bank (20-ES-0021). Ethical approval for the GS:SFHS study was obtained from the Tayside Committee on Medical Research Ethics (on behalf of the National Health Service). Ethical approval for the GS:21CGH study was obtained from the Scotland A Research Ethics Committee. Patients/participants provided their written informed consent to participate in this study.

11. Acknowledgements

We thank the Generation Scotland participants and the Generation Scotland team. Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006]. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland, was funded by the Medical Research Council UK and has been and is currently funded by the Wellcome Trust (STratifying Resilience and Depression Longitudinally (STRADL) Reference 104036/Z/14/Z and Generation Scotland: NextGenScot Reference 216767/Z/19/Z).

We also thank the annotators (Liam Lee, Michael Walsh, Freya Pellie, Karen Ferguson and William Whiteley) who contributed to creating the GS-BrainText gold standard labels.

Authors (initials in brackets) were supported by:

- Turing Fellowships (B.A. and C.G.) and a Turing project (B.A. and A.C.) funded by the Alan Turing Institute (EPSRC grant EP/N510129/1)
- Legal& General Group as part of the Advanced Care Research Centre (B.A. and A.C.)
- The AIM-CISC project funded by the National Institute for Health Research (NIHR202639; B.A.)
- An MRC Clinician Scientist Award (G0902303; W.W.)
- A Scottish Senior Clinical Fellowship (CAF/17/01; W.W.) The funders had no role in the conduct of the study, interpretation or the decision to submit for publication. The views expressed are those of the authors and not necessarily those of the funders.

12. Bibliographical References

Beatrice Alex, Claire Grover, Richard Tobin, Cathie Sudlow, Grant Mair, and William Whiteley. 2019. Text mining brain imaging reports. *Journal of Biomedical Semantics*, 10(Supplement 1):23.

Michael PJ Camilleri, Dorian Gouzou, Salim Al-Wasity, Muthu RK Mookiah, María Valdes Hernandez, Bea Alex, Sotirios A Tsaftaris, Andrew Brooks, Ruairidh MacLeod, Honghan Wu, et al. 2025. A large dataset of brain imaging linked to health systems data: the curation and access to a whole system national cohort from NHS Scotland. *medRxiv*, pages 2025–10.

Arlene Casey, Emma Davidson, Claire Grover, Richard Tobin, Andreas Grivas, Huayu Zhang, Patrick Schrempf, Alison Q. O'Neil, Liam Lee, Michael Walsh, Freya Pellie, Karen Ferguson, Vera Cvoro, Honghan Wu, Heather Whalley, Grant Mair, William Whiteley, and Beatrice Alex. 2023. Understanding the performance and reliability of NLP tools: a comparison of four NLP tools predicting stroke phenotypes in radiology reports. *Frontiers in Digital Health*, 5.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Sunyang Fu, Liwei Y Leung, Anne-Olivia Raulli, David F Kallmes, Kerry A Kinsman, Kristin B Nelson, Thomas K Clark, Patrick H Luetmer, Peter R Kingsbury, David M Kent, et al. 2020. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. *BMC Medical Informatics and Decision Making*, 20(1):60.

Sunyang Fu, Liwei Y Leung, Yiqing Wang, Anne-Olivia Raulli, David F Kallmes, Kerry A Kinsman, Kristin B Nelson, Thomas K Clark, Patrick H Luetmer, Peter R Kingsbury, et al. 2019. Natural language processing for the identification of silent brain infarcts from neuroimaging reports. *JMIR Medical Informatics*, 7(2):e12109.

Andreas Grivas, Beatrice Alex, Claire Grover, Richard Tobin, and William Whiteley. 2020. Not a cute stroke: analysis of rule-and neural network-based information extraction systems for brain radiology reports. In *The 11th International Workshop on Health Text Mining and Information Analysis at EMNLP 2020*, pages 24–37. Association for Computational Linguistics.

Matthew H Iveson, Mome Mukerjee, Emma M Davidson, Huayu Zhang, Laura Sherlock, Emily L Ball, Grant Mair, Alice Hosking, Heather Whalley, Michael T C Poon, Joanna M Wardlaw, David Kent, Richard Tobin, Claire Grover, Beatrice Alex, and William N Whiteley. 2025. Clinically reported covert cerebrovascular disease and risk of neurological disease: a whole-population cohort of 395,273 people using natural language processing. *medRxiv*.

Caroline Jackson, Laura Crossland, Michele Dennis, Joanna Wardlaw, and Cathie Sudlow. 2008. Assessing the impact of the requirement for explicit consent in a hospital-based stroke study. *QJM: Monthly Journal of the Association of Physicians*, 101(4):281–289.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Charlotte Maschke, Peter Hadar, Yicheng Zhang, Jian Li, Gauri Ganjoo, Andrew Hoopes, Alessandro Guazzo, Aditya Gupta, Manohar Ghanta, Bruce Nearing, Christine Tsien Silvers, Bharath Gunapati, Robert Thomas, Jennifer A. Kim, Shibani S. Mukerji, Adrian Dalca, Sahar Zafar, Alice D. Lam, Emmanuel Mignot, and M Brandon Westover. 2025. The brain imaging and neurophysiology database: Binding multimodal neural data into a large-scale repository. *medRxiv*.

Hannah Milbourn, Daniel McCartney, Anne Richmond, Archie Campbell, Robin Flaig, Sarah Robertson, Chloe Fawns-Ritchie, Caroline Hayward, Riccardo E Marioni, Andrew M McIntosh, et al. 2024. Generation scotland: an update on scotland's longitudinal family health study. *BMJ open*, 14(6):e084719.

Kristiina Rannikmäe, Honghan Wu, Sophie Tominey, William Whiteley, Naomi Allen, and Cathie UK Biobank Stroke Study Sudlow. 2021. Developing automated methods for disease subtyping in UK Biobank: An exemplar study on stroke. *BMC Medical Informatics and Decision Making*, 21(191).

Patrick Schrempf, Hannah Watson, Shadia Mikhael, Maciej Pajak, Matúš Falis, Aneta Lisowska, Keith W Muir, David Harris-Birtill, and Alison Q O'Neil. 2020. Paying per-label attention for multi-label extraction from radiology reports. In *International Workshop on Interpretability of Machine Intelligence in Medical Image Computing*, pages 277–289. Springer.

Patrick Schrempf, Hannah Watson, Eunsoo Park, Maciej Pajak, Hamish MacKinnon, Keith W Muir, David Harris-Birtill, and Alison Q O'Neil. 2021. Templated text synthesis for expert-guided multi-label extraction from radiology reports. *Machine Learning and Knowledge Extraction*, 3(2):299–317.

Blair H Smith, Archie Campbell, Pamela Linksted, Bridie Fitzpatrick, Cathy Jackson, Shona M Kerr, Ian J Deary, Donald J MacIntyre, Harry Campbell, Mark McGilchrist, Lynne J Hocking, Lucy Wisely, Ian Ford, Robert S Lindsay, Robin Morton, Colin N A Palmer, Anna F Dominiczak, David J Porteous, and Andrew D Morris. 2012. Cohort profile: Generation scotland: Scottish family health study (gs:sfhs). the study, its participants and their potential for genetic research on health and illness. *International Journal of Epidemiology*, 42(3):689–700.

Blair H Smith, Harry Campbell, Douglas Blackwood, John Connell, Mike Connor, Ian J Deary, Anna F Dominiczak, Bridie Fitzpatrick, Ian Ford, Cathy Jackson, et al. 2006. Generation scotland: the scottish family health study; a new resource for researching genes and heritability. *BMC medical genetics*, 7(1):74.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Dominic Sykes, Andreas Grivas, Claire Grover, Richard Tobin, Cathie Sudlow, William Whiteley, Andrew McIntosh, Heather Whalley, and Beatrice Alex. 2021. Comparison of rule-based and neural network models for negation detection in radiology reports. *Natural Language Engineering*, 27(2):203–224.

Emily Wheater, Grant Mair, Cathie Sudlow, Beatrice Alex, Claire Grover, and William Whiteley. 2019. A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records. *BMC Medical Informatics and Decision Making*, 19(1):184.

SASTA Self Assessment: An efficient human-in-the-loop strategy for developmental and pathological language analysis

Jan Odijk, Jelte van Boheemen, Xander Vertegaal, Tessel Boerma, Marijn Schraagen
Utrecht University, the Netherlands
{j.odijk, j.vanboheemen, a.j.j.vertegaal, t.d.boerma, m.p.schraagen}@uu.nl

Abstract

This paper introduces SASTA self-assessment, i.e., a self-assessment procedure for the SASTA application for semiautomatic analysis of spontaneous language transcripts for Dutch. We introduce SASTA and the methods that it supports. These methods are used to assess the language development of young children and to assess the language skills of patients with aphasia. We illustrate this with typical example utterances. The performance of SASTA is good but not good enough for fully automatic use. The self-assessment procedure attempts to automatically identify utterances that require revision by a human expert. The self-assessment procedure gives promising results for datasets for the ASTA and STAP methods. Significant improvements are still needed for the TARSP method, but there is still potential for such improvements, and we sketch some directions to achieve such improvements.

Keywords: language acquisition, developmental language disorders, aphasia, parsing, grammar correction, self assessment, Dutch

1. Introduction

This paper introduces SASTA self-assessment, i.e., a self-assessment procedure for the SASTA application for semiautomatic analysis of spontaneous language transcripts for Dutch. We introduce SASTA (Section 2) and the methods that it supports. These methods use language measures to assess the language development of young children and to assess the language skills of patients with aphasia (Section 2.1). We illustrate this with typical example utterances in Section 2.2. The performance of SASTA (reported in Section 2.3) is good but not good enough for fully automatic use. We introduce a self-assessment procedure that attempts to automatically identify utterances that require revision by a human expert. We set an initial target for the F1 score that should be achieved for a sample after self-assessment and revision of the selected utterance by a human expert (Section 3) and observe that this target is in principle reachable for all samples from all datasets that we use in this research and for all methods. We describe the self-assessment procedure in more detail in Section 4, identifying two classes of criteria that can be used for this purpose. The results of the self-assessment procedure are reported in Section 5, both for the training data and for the test data. We discuss some relevant related work in Section 6. Section 7 summarises our findings and suggests concrete work and some directions that may lead to improvements of the self-assessment procedure.

2. SASTA

SASTA[1] (Odijk, 2020; Odijk et al., 2025) is an application for semiautomatic analysis of spontaneous language transcripts for Dutch. It supports established methods for the analysis of such transcripts, in particular TARSP (aimed at children 1-4 years of age) (Schlichting, 2005, 2017), STAP (aimed at children 4-6 years of age) (van Ierland et al., 2008; Verbeek et al., 2007) and ASTA (aimed at patients with aphasia) (Boxum et al., 2013).

SASTA automates the analysis of samples in accordance with these methods, with a focus on grammatical analysis, and a partial analysis of grammatical errors.

The application is continuously being improved but it is already used to support research into language acquisition, developmental language disorders, and aphasia at various universities, and as a tool to support diagnosis and treatment in several clinics.

2.1. Language Measures

The methods that SASTA supports provide recipes for analysing the transcript of a language sample. A sample consists of 35-50 utterances, with the exact size depending on the method. The methods define *language measures*, i.e counts of how often specific lexical and grammatical phenomena occur in the sample:

- # specific function words and function and content word classes:

[1] SASTA is the acronym for the Dutch version of Semi-automatic Analysis of Spontaneous Language.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 103–112
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

  - # nouns, # auxiliary verbs, ...
- # specific morphological properties:
  - verbal inflection, noun plurals, diminutives, compounds, ...
- # specific syntactic constructions
  - various clause structures, NP structures, PP structures, ...
- (for some) # specific grammar error classes
  - subject verb agreement, determiner-noun agreement, overgeneralised verb forms, ...

The methods assign a code for each occurrence of these phenomena. The counts and some aggregations on the counts are summarised in a method-specific *language profile* form. The results are compared to normative results, which (often in combination with results from other tests) is used to make an assessment of the language skills and to set up a treatment plan. The importance of language sample analysis in clinical practice is emphasised by Klatte et al. (2022). Until recently, analysis of such samples had to be done fully manually. SASTA automates this analysis.

2.2. Example Utterances

SASTA takes as input a language sample, which is a transcript of a recording of a conversation between the patient or child, and an interviewer. It consists of a sequence of utterances, together with metadata on the transcript and with metadata on individual utterances (at least the speaker). These utterances contain a lot of deviant language because the speakers are very young children who are still acquiring language, some of which may have a language disorder, or patients with aphasia. The transcribers usually transcribe faithfully what the speakers say, which results in utterances with many words that do not exist in Dutch or must be interpreted in an unusual way. Some examples of such utterances are in (1-3). For each example, we provide a gloss in which we try to reflect any deviant aspects of the utterance, the utterance as corrected by SASTA and its gloss, and a translation:

(1) TARSP examples

a. moe peelkaal toe
mus pay-all to

moet naar speelzaal toe
must to play-hall to

'(I) must go to the playing hall'

b. nu ga ik dees [: deze] pobreren
now go I thes [: this.one] ty
[: proberen].
[: try]

Nu ga ik deze proberen.
now go I this.one try

'Now I will try this one'

(2) STAP examples

a. die si o(p) wate(r).
that zi o(n) wate(r)

die zit op water
that sits on water

'That one sits on the water'

b. weet je waar ik woont?
know you where I lives?

weet je waar ik woon?
know you here I live?

'Do you know where I live?'

(3) ASTA examples

a. Op euh eventien zeventien okt ja
On eh eventeen seventeen Oct yeah
oktober is onze ja
October is our yeah

Op zeventien oktober is onze
On seventeen October is our

'On October 17th is our'

b. dan heeft het allemaal gebeurd
then gets it all happened

toen is het allemaal gebeurd
then is it all happened

'It all happened then'

The examples illustrate various kinds of deviancy, e.g., omitted sounds and syllables (*peelkaal* instead of ***s**peel**lo**kaal*), and informal spoken language variants (*moe* instead of *moet*) in (1a), wrong pronunciations (*pob**r**eren* instead of *p**r**oberen*) in (1b), agreement errors (*ik woon**t*** instead of *ik woon*) in (2b), wrong use of auxiliaries (*heeft* instead of *is* in (3b)), and filled pauses (*euh*), hesitations (*ja*) and false starts (*eventien*, *okt*) in an incomplete utterance (3a).

Transcribers can use inline CHAT annotations (MacWhinney, 2000) to clarify the transcript (e.g., *pobreren [: proberen]*, and *o(p)*), but these are used inconsistently (e.g. *o(p)* but not *si [: zit]* for *si* in (2a)) even by individual transcribers.

Table 1 provides the codes that have to be assigned to these utterances.

SASTA analyses all utterances by the target speaker in the sample. For each target speaker utterance, it attempts to detect any deviancy. If it detects any deviancy, it generates alternatives,

Method	Example	Codes
TARSP	moe peelkaal toe	BX, SamZn
TARSP	nu ga ik dees [: deze] pobreren [: proberen].	Avn,Hww i,Inv,OndWBVC,ik
STAP	die si o(p) wate(r).	BB p,N,PV,zelfst.vnw.
STAP	weet je waar ik woont?	OS,PV,PV
ASTA	Op euh eventien zeventien okt ja oktober is onze ja	MLUX,MLUX,MLUX,N,PV,SSZX, SSZX, lemma=oktober
ASTA	dan heeft het allemaal gebeurd	LEX,PV,lemma=gebeuren

Table 1: Example utterances with method and applicable codes

For example, in the utterance *moe peelkaal toe* it detects the following deviant or potentially deviant aspects:

- the word *moe* is an existing Dutch word (meaning *tired*) but it is also an informal pronunciation of the word *moet* (with meaning *must*). It therefore generates *moet* as an alternative to *moe*.
- the word *peelkaal* is not an existing Dutch word. SASTA generates an alternative using a dedicated child-language specific spelling corrector which has been developed as a component of SASTA. This generates the word *speelzaal* as an alternative.
- the word *toe* is an existing Dutch word. It is highly ambiguous: it can be an interjection (meaning 'please'), in certain regions an adjective (meaning *closed*), and it can be a postposition (meaning 'to'). The postposition *toe* never occurs on its own in the standard Dutch language (it usually occurs in the combination *naar ... toe*), but young children often use it on its own. SASTA adds the preposition *naar* before *peelkaal* as an alternative.

This results in 8 variants: all combinations of *moe* or *moet*, *peelkaal* or *speelzaal*, and presence or absence of *naar*. Each correction that is applied is recorded in correction metadata. SASTA parses all of these 8 variants (using the Alpino parser (Bouma et al., 2001; van Noord, 2006)). Then it evaluates the 8 parses using *parse evaluation criteria* and selects the utterance with the best parse.

There are currently around 30 parse evaluation criteria. Some are very generic, others are highly specific for particular grammatical configurations. An example of a parse evaluation criterion is the following: if the parser cannot analyze the utterance as a single utterance, it attempts to analyze it as a sequence of utterances. Each utterance in such a sequence gets the grammatical relation *dp* (discourse part). Since the string offered is a single utterance, presence of the grammatical relation *dp* in the parse tree is an indication that the parser could not analyze this utterance correctly. Such utterances will be less preferred when selecting the utterance with the best parse. We will discuss some more parse evaluation criteria in section 4.

The words in the best parse are then replaced again with the original words, and inserted words are deleted, by a procedure called *backplacement*. The result is that the utterance *moe peelkaal toe* has the parse tree of *moet naar speelzaal toe*, so *moe* is analysed as a finite modal verb, and *peelkaal toe* as an adverbial modifier.

The language measures have been implemented as queries on parse trees: these queries are applied to the original utterance with the selected best parse. The results of these queries are used to generate an annotated transcript and a digital version of the method-specific language profile form. For example (1a), it generates the codes *BX* (adverbial modifier and one other constituent) for the whole utterance, and *SamZn* (compound noun) for *peelkaal*,[2] exactly as required.

SASTA is, of course, not perfect. That is why SASTA has the term *semiautomatic* in its name. The user can edit the annotated transcript and resubmit it to SASTA, after which an updated language profile form is generated.

The SASTA application is publicly available[3], the source code is open software[4] and available on GitHub.[5]

2.3. Performance of SASTA

SASTA is being developed on the basis of a range of datasets. Some of these datasets are used for training, i.e. they are inspected to see what new alternatives must be generated, whether queries

[2]The backplacement procedure contains a specific procedure to determine whether the actually pronounced word (*peelkaal*) must be considered a compound if its correction (*speelzaal*) is one.

[3]https://sasta.hum.uu.nl/.

[4]BSD 3-Clause License, https://opensource.org/license/bsd-3-clause.

[5]https://github.com/CentreForDigitalHumanities/sasta, https://github.com/UUDigitalHumanitieslab/sastadev and https://github.com/CentreForDigitalHumanities/sasta-scripts.

that implement the language measures have to be adapted, etc. Other data sets are only used to test the system.

The performance of SASTA is measured in terms of recall (how many of the codes in the reference are actually generated by SASTA), precision (how many of the codes generated by SASTA occur in the reference), and by F1 score (the harmonic mean of recall and precision). Here we will discuss F1 scores only.

The performance of SASTA in terms of F1 scores is given in Table 2.

One can see that the performance is good with the F1 score on average between 88.5% and 94.4%. However, individual samples sometimes have a much lower score, with a minimum of 77.6%.

This level of performance still requires manual correction of the sample analysis to achieve the desired level of performance. If the whole sample has to be checked manually, the time gain by using SASTA is drastically reduced.[6] That is where SASTA self-assessment comes into play.

3. Self Assessment

In order to avoid that the whole sample has to be reviewed manually to check whether SASTA assigned the right codes, it would be ideal if SASTA could identify the utterances for which it thinks or suspects that its analysis is wrong, i.e., we need a form of SASTA Self-Assessment. That is the main topic of this paper.

The idea is that SASTA identifies a limited number of utterances that have to be manually checked. We assume that manually checking these utterances will lead to a 100% F1 score for these utterances.

Of course, we attempt to achieve as high an F1 score as possible, with as few revisions as possible, but as an initial (somewhat arbitrary) target, we aim to achieve a 95% F1 score for each sample by revising maximally 15 utterances per sample. 15 utterances is less than half of the utterances for an ASTA sample, 37.5% of the utterances for a TARSP sample and 30% of the utterances for a STAP sample.

Since we have expert-annotated reference data containing language measure codes for conversation transcripts, we can determine what this approach could achieve in the best possible case. This has been represented in Table 3. Note that this reference will change when SASTA is modified. The table represents the reference as determined on 2025-10-20.

One can see that all samples in all datasets reach an F1 score of at least 95% by revising at most 14 utterances. The average number of utterances that have to be modified over all samples in all datasets is 4.2 utterances per sample. Of course, this can only be achieved if the right utterances are selected for correction. For this selection, one needs assessment criteria.

4. Assessment Criteria

We identified two types of criteria that can be used for self-assessment: parse assessment criteria, and code assessment criteria. We will discuss each type in a separate subsection. In Section 4.3 we describe the strategy for testing these criteria against the training data and extending them.

4.1. Parse assessment criteria

If the parse of an utterance is wrong, it is very likely that the application of the queries that implement the language measures yields wrong results. For this reason, we assessed the parse evaluation criteria for their potential as a criterion for self-assessment. We identified 15 such criteria as having potential, and 17 as having no or low potential. These criteria determine the number of times that a specific configuration occurs in the parse tree. Some criteria are negative (so the lower the number of occurrences, the better), others are positive (so the higher the number of occurrences, the better).

We list some of the more generic assessment criteria:

dp # *dp* relations that the parse tree contains

unknown words # 'unknown' words that the parse tree contains[7]

extra-grammatical # content words that have no grammatical connection to the other words in the parse tree

Note that some parse evaluation criteria have to be adapted to function as a self-assessment criterion. In example (1a) discussed earlier, the utterance has the parse tree of *moet naar speelzaal toe* but it contains the words *moe peelkaal toe*, which includes the 'unknown' word *peelkaal*. The system only considers a word such as *peelkaal* as 'unknown' if it was not replaced with a known word

[6] SASTA is still useful even if the whole sample has to be checked, because SASTA increases recall (human annotators often overlook items that have to be coded) and because reviewing assigned codes is faster than assigning codes from scratch.

[7] We operationalised the notion of 'unknown word' by using some external lexicons, taking into account some very productive morphological processes.

Dataset	Use	Method	Avg F1	Min F1	Max F1
auristrain	train	tarsp	92.7	85.5	96.9
vkltarsp	train	tarsp	93.5	87.4	97.8
vklstap	train	stap	93.8	90.9	95.8
vklasta	train	asta	88.5	77.6	94.7
auristest	test	tarsp	92.6	83.5	97.3
elsdejong	test	stap	92.2	87.4	95.3
vklstap2	test	stap	93.3	90.2	96.2
vklasta2	test	asta	89.4	83.6	94.0
wrijpma	test	asta	94.9	90.6	98.2

Table 2: SASTA performance in terms of F1 score (measured 2025-10-20)

Use	Dataset	Method	#	$\geq$ target	Or. Avg F1	Avg F1	Min F1	Max F1	Avg #Utt	Min #Utt	Max # Utt
train	auristrain	tarsp	28	28	92.5	95.8	95.0	97.6	4.5	1	7
train	vkltarsp	tarsp	11	11	93.4	96.0	95.0	98.5	2.2	1	6
train	vklstap	stap	9	9	93.9	95.7	95.0	96.5	2.3	1	6
train	vklasta	asta	10	10	86.9	95.4	95.0	95.9	6.6	1	14
test	auristest	tarsp	8	8	91.9	96.0	95.4	98.4	3.0	1	7
test	vklstap2	stap	13	13	93.4	95.5	95.0	96.7	3.2	1	7
test	elsdejong	stap	10	10	92.2	95.4	95.0	96.1	4.1	1	10
test	vklasta2	asta	6	6	88.3	95.4	95.0	95.7	5.8	1	9
test	wrijpma	asta	25	25	94.8	96.2	95.2	98.6	1.4	1	4

Table 3: SASTA Self Assessment Reference (measured 2025-10-20), The table specifies for each dataset its use (for training or for testing), the method, the number of samples (in the column with header **#**), the number of samples for which the F1 target is reached, the average F1 score for each dataset (header **Or. Avg F1**), the average, minimal and maximal F1 score after correcting the relevant utterances, and the average, minimal and maximal number of utterances that must be revised.

by the correction mechanism: it can determine this by inspecting the correction metadata.

Currently, we have identified around 30 assessment criteria of this type. Table 4 provides some examples with a brief explanation. This table contains only verbal descriptions, but each criterion has been implemented as a query on the parse tree and/or the codes generated. Note that SASTA, using these criteria, can determine that the parse tree is very likely incorrect, but it cannot determine what is the correct one. Furthermore, many configurations that the criteria describe can be correct in special cases: what is important is that these criteria identify as many utterances with truly incorrect parses and identify as little as possible utterances with correct parses. Two different criteria can have overlapping results: in a next version of the system we will try to reduce any superfluous redundancy.

4.2. Code assessment criteria

A second type of criterion for self-assessment involves the queries that are used to implement the language measures. For example, under the TARSP-method, if an utterance contains a verb and at least one other real word (i.e., not interpunction, filled pause or interjection), then a code from a certain subset must have been assigned. If no such code is present, the analysis is very likely wrong. The cause for this can be a wrong parse, but it could also be an error or omission in the query. If the cause is in the query, we of course try to improve the query, but that is not always possible: criteria that identify utterances in which this occurs can be used as self-assessment criteria. We currently have three such criteria.

4.3. Developing criteria

We applied an initial set of assessment criteria to the various training datasets and analysed the mismatches, in particular the utterances that were not identified by the self-assessment criteria but that should have been. The analysis of these mismatches revealed that each case can be classified in one of three types:

1. the manual annotation in the reference contains an error. This actually happened quite often. Some of these errors were real errors in the reference data, but most were caused because for certain types of constructions specific conventions for annotating them were fixed during the development of SASTA. The

Short Description	Explanation
Bad categories	# categories that occur mostly if the utterance was parsed incorrectly
Basic replacement not applied	For a list of frequent incorrect or informal pronunciations replacements are generated as an alternative. But if such a word does not get replaced, the parse will very likely be wrong
Non-neuter determiner with neuter noun	This concerns an agreement error that can sometimes be corrected by SASTA, but if it is not corrected, the parse is very likely wrong
Infinitive probably wrongly parsed as subject	Infinitives as a subject are very rare in young children's language
Multiple main clauses	The presence of multiple main clauses makes it very likely that the utterance has been parsed incorrectly
Word unknown to the parser	Alpino attempts to guess the properties of words that it does not know. It does this often successfully, but not always (e.g. for the word *peelkaal* Alpino guessed that it is an adjective, which is understandable since it ends in *-aal*, a very common derivational suffix to form adjectives, but in utterance (1a) this is not correct.
Likely wrong indirect object	In certain configurations indirect objects are unlikely
Mismatching conjunct	The Alpino parser is (like all parsers) bad at analysing conjuncts correctly; if the conjuncts have a mismatch in category, it is very likely that the parse is wrong
Likely wrong analysis as locative and nominal predicate	The combination of a locative and a nominal predicate usually means that the utterance has been parsed incorrectly
Direct object with a (preferably) intransitive verb	If a direct object co-occurs with a verb that is (preferably) intransitive, it is highly likely that the parse is wrong
Wrong part of speech	Certain words are ambiguous (e.g., as noun or as verb), but one of these readings is very unlikely to occur in the samples. For example, the word *kijk* 'look' can be a verb or a noun, but it is very unlikely that *kijk* occurs as a noun in young children's language samples
Semantic incompatibility	An argument of a verb is semanticallly incompatible with the semantic selection properties of this verb for that argument.
Ellipsis applied incorrectly	Alpino is (like all parsers) not very good at analysing cases of ellipsis. This criterion selects the most likely wrong parses of ellipted words and phrases

Table 4: Example assessment criteria with a brief explanation

original reference data were made without these conventions in place, so were inconsistent with regard to them. We discuss such inconsistencies with the original data providers, and adapt the reference in accordance with the annotation convention agreed upon.[8]

2. the criteria did not cover them: in these cases we improved the criteria or added new criteria

3. the utterance was correctly not selected but the SASTA queries contain an error or omission: in these cases we try to improve these queries, and where we do not succeed we try to add new criteria.

[8]This concerns, for example, decisions on certain cases that could be analysed as a complement or as a modifier; which sequences of two adverbs form one constituent, and which sequences form two constituents, and similar cases.

5. Self Assessment Results

The results of self-assessment for the training data as measured on 2025-10-20 have been summarised in Table 5.

SASTA produces an F1 score $\geq$ 95% without SASTA self-assessment for some samples. These are included in the results, but they are few, as is clear from Table 6.

We observe that the results for datasets *vklstap* and *vklasta* with 8 out of 9 and 10 out of 10 samples reaching the target are good. The average number of utterances per sample that have to be revised (10.3 and 13.4) is much higher than the ideal case (2.2 and 6.6), but we did not optimise for this yet. The results for the TARSP datasets *auristrain* and *vkltarsp* are more mixed, with 20 out of 28 and 9 out of 11 samples reaching the target set. The criteria used are clearly not able to select all the right utterances for revision yet. These require

dataset	method	#	OK	Orig. Avg f1	Avg F1	Min F1	Max F1	Avg #Utt	Min #Utt	Max #Utt	Min f1 Δ	Max f1 Δ	Avg f1 Δ
auristrain	tarsp	28	20	89.9	93.9	88.2	99.7	6.0	2	12	0.0	6.9	3.2
vkltarsp	tarsp	11	9	93.4	96.1	91.4	98.8	6.1	2	15	0.5	8.9	2.8
vklstap	stap	9	8	93.9	96.3	94.3	97.6	10.3	6	15	0.8	4.3	2.5
vklasta	asta	10	10	86.9	96.1	89.0	98.8	13.4	8	15	2.0	18.7	10.5

Table 5: Self-assessment results for the training data (2025-10-20).The table provides columns for the dataset, the method, the number of samples (#), the number of samples for which the target of 95% F1 score is reached (OK), the original average F1 score, the average, minimal and maximal F1 score after application of self-assessment, the average, minimal and maximal number of utterances that have to be revised, and the minimal, maximal and average increase in F1 score after self-assessment.

use	dataset	method	#	# $\geq$ 95%
train	auristrain	TARSP	28	5
train	vkltarsp	TARSP	11	3
train	vklstap	STAP	9	4
train	vklasta	ASTA	10	0
test	auristest	TARSP	8	2
test	vklstap2	STAP	13	2
test	vklasta2	ASTA	6	0
test	elsdejong	STAP	11	1
test	wrijpma	ASTA	25	11

Table 6: Number of samples per dataset and the number of samples with an F1 score $\geq$ 95% by SASTA without self-assessment per dataset (2025-10-20)

more work.

The results of self-assessment for the test data as measured on 2025-10-20 have been summarised in Table 7. This table has the same structure as Table 5.

The results for the STAP data *vklstap2*, with 11 out of 13 is promising, but the results for the dataset *elsdejong*), with 6 out of 11 samples reaching the 95% F1 target still requires improvement. The ASTA dataset *wrijpma* shows a perfect score with 25 out of 25 samples reaching the target, but the original average F1 score for samples in this dataset was very high to begin with (94.8%), as is the number of samples that start with an F1 score $\geq$ 95% (11 out of 25). It is very likely that SASTA self-assessment can reach a higher target for this dataset. In the *vklasta2* dataset 3 out of 6 samples reach the target, so an improvement is clearly desirable here.

The *auristest* data, with 5 out of 8 samples reaching the target, also still require improvement. The average number of utterances to be revised is quite low (5.1), which suggests that the current selection criteria are not yet able to select the right utterances for revision.

6. Related Work

The focus of this paper is on SASTA self-assessment. For this reason we will not discuss work that relates to automating methods to assess the language skills of young children and patients with aphasia. For this , we refer to (Odijk, 2020).

The code assessment criteria are, to our knowledge, unique, though they are instantiations of a general technique, viz. formulating a less specific query to determine whether a code from a specific subset must be present, raise an error if this is not the case, and suggest the members of the subset as candidate codes.

The other method used to do self-assessment is by attempting to automatically detect parsing errors. This makes a lot of research dealing with this topic or with automatically detecting annotation errors in treebanks relevant for this paper.

Ambati et al. (2010) created a tool to detect parsing and annotation errors in Hindi treebanks The tool contains rules implementing checks on annotation conventions and statistical checks on local tree configurations. For a different project, one of the authors made a similar tool for Dutch (Odijk et al., 2018),[9] with the statistics obtained from the manually verified Spoken Dutch Corpus (Oostdijk et al., 2002) and Lassy treebanks (van Noord et al., 2013).

Many researchers, inter alia (van Halteren, 2000; Boyd et al., 2008; Dickinson and Meurers, 2003, 2005; Dickinson, 2011; Dickinson and Smith, 2011; Krasnowska and Przepiórkowski, 2013) use variation in annotations as a basis for automatically

[9] https://github.com/JanOdijk/ace

dataset	meth	#	OK	Orig. Avg F1	Avg F1	Min F1	Max F1	Avg #Utt	Min #Utt	Max #Utt	Min F1 Δ	Max F1 Δ	Avg F1 Δ
auristest	tarsp	8	5	91.9	94.9	91.6	97.8	5.1	2	8	0.6	8.1	2.9
vklstap2	stap	13	11	93.4	96.4	93.3	97.8	12.5	6	15	1.0	7.4	3.2
elsdejong	stap	11	6	91.8	95.8	93.7	98.6	13.5	9	15	0.9	7.6	4.5
vklasta2	asta	6	3	88.4	95.1	91.6	98.2	14.7	13	15	3.1	13.6	8.0
wrijpma	asta	25	25	94.8	97.9	95.9	100.0	12.6	6	15	1.1	6.3	3.2

Table 7: Self-assessment results for the test data. (2025-10-20)

detecting likely errors.

The work on error mining in treebanks to identify elements that are the cause for wrongly parsed sentences is clearly also relevant (van Noord, 2004; Sagot and de la Clergerie, 2006; de Kok et al., 2009; Suzuki et al., 2016). This research is usually aimed at improving a treebank and/or a parser. These works attempt to determine the 'suspicion' of elements (e.g. words) or combinations of elements (e.g., n-grams) based on their frequencies in wrongly parsed sentences in comparison to their frequencies in correctly parsed sentences. Some of these methods work iteratively, e.g. running the same procedure on a continuously improving treebank until the suspicions stabilize.

Volokh and Neumann (2011) parse a manually verified treebank with two parsers trained on the treebank and use the differences between these results as likely sources of annotation errors, and they even try to correct these automatically. (Agrawal et al., 2013) also use a parser to identify errors in a manually verified treebank, and they emphasize that using a parser will identify inconsistencies in human annotation.

These methods are usually applied to find errors in large treebanks that contain enough data to work with frequencies and derivatives of frequencies, or to train parsers on. In our research, the total amount of data (training and testing) is relatively small (currently 2300 utterances for training, 2600 for testing), so that it is not obvious that these methods can be usefully applied. (cf. (van Noord, 2004): 'this technique obtains very good results if it is applied to large sets of sentences'.

For this reason we have so far used a rule-based approach to identify likely parsing errors. In a future version, we will also try to use statistics obtained from the manually verified Van Kampen CHILDES treebank (Odijk et al., 2018) to detect likely parse errors, at least for the data involving young children. This treebank contains 109k utterances, 47k of which are utterances made by young children.

7. Conclusions and Future Work

We have introduced a mechanism of self-assessment for SASTA. The current version of the self-assessment procedure yields promising results for ASTA and STAP datasets, but clearly still requires improvement for the TARSP datasets.

We introduced two types of assessment criteria: parse assessment criteria, which attempt to identify wrong parses, and code assessment criteria, which attempt to identify incorrectly missing codes.

The work on SASTA self-assessment is still ongoing, and there is still potential for further improvement. We finished the analysis of the utterances that were incorrectly not selected but we did not implement all new criteria or SASTA improvements yet. We also aim to investigate some other criteria, in particular whether we can somehow use confidence values for language measures in new criteria.

We will also try to make use of the statistics obtained from the manually verified Van Kampen CHILDES treebank (Odijk et al., 2018) to detect likely parse errors, at least for the data involving young children.

We also want to analyse the utterances that have incorrectly been selected, to try to reduce their number as much as possible. Though some criteria are already specific for a particular method, this might be necessary for other criteria that are currently applied for all methods as well.

We have inventoried a number of criteria that can be used to sort the selected utterances, so that the utterances that increase the F1 score most come earlier. Examples of such criteria are the utterance length relative to the number of codes assigned and the average F1 increase in the training data. However, we have not experimented with these criteria yet.

In short, there is still a lot of potential for improvement, and we hope to report on better results of new versions of the self-assessment procedure in the near future.

Acknowledgements

The research reported on in this paper was partially financed by the Utrecht University Human-Centered AI focus area. We thank the anonymous reviewers for their comments.

8. Bibliographical References

Bhasha Agrawal, Rahul Agarwal, Samar Husain, and Dipti M. Sharma. 2013. An automatic approach to treebank error detection using a dependency parser. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing. CICLing 2013. Lecture Notes in Computer Science, vol. 7816*, pages 294–303. Springer, Berlin, Heidelberg.

Bharat Ram Ambati, Mridul Gupta, Samar Husain, and Dipti Misra Sharma. 2010. A high recall error identification tool for Hindi treebank validation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, 37(1):45–59.

Elsbeth Boxum, Fennetta van der Scheer, and Mariëlle Zwaga. 2013. *Analyse voor Spontane Taal bij Afasie. Standaard in samenwerking met de VKL*. VKL. https://klinischelinguistiek.nl/uploads/201307asta4eversie.pdf.

Adriane Boyd, Markus Dickinson, and W. Detmar Meurers. 2008. On detecting errors in dependency treebanks. *Research on Language and Computation*, 6(2):113–137.

Daniël de Kok, Jianqiang Ma, and Gertjan van Noord. 2009. A generalized method for iterative error mining in parsing results. In *Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks (GEAF 2009)*, pages 71–79, Suntec, Singapore. Association for Computational Linguistics.

Markus Dickinson. 2011. Detecting ad hoc rules for treebank development. *Linguistic Issues in Language Technology (LiLT)*, 4(3).

Markus Dickinson and W. Detmar Meurers. 2003. Detecting inconsistencies in treebanks. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, pages 45–56, Växjö, Sweden. Växjö University Press.

Markus Dickinson and W. Detmar Meurers. 2005. Detecting errors in discontinuous structural annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 322–329, Ann Arbor, Michigan. Association for Computational Linguistics.

Markus Dickinson and Amber Smith. 2011. Detecting dependency parse errors with minimal resources. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 241–252, Dublin, Ireland. Association for Computational Linguistics.

Inge Klatte, Vera Van Heugten, Rob Zwitserlood, and Ellen Gerrits. 2022. Language sample analysis in clinical practice: Speech-language pathologists' barriers, facilitators, and needs. *Language, Speech, and Hearing Services in Schools*, 53:1–16.

Katarzyna Krasnowska and Adam Przepiórkowski. 2013. Detecting syntactic errors in dependency treebanks for morphosyntactically rich languages. In *Language Processing and Intelligent Information Systems – 20th International Conference, IIS 2013 (selected papers)*, pages 69–79. Springer.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3 edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Jan Odijk. 2020. Towards semi-automatic analysis of spontaneous language for Dutch. In *Selected papers from the CLARIN Annual Conference 2020*, volume 180 of *Linköping Electronic Conference Proceedings*, pages 165–175. Linköping University Press, Linköping, Sweden.

Jan Odijk, Alexis Dimitriadis, Martijn Van der Klis, Marjo Van Koppen, Meie Otten, and Remco Van der Veen. 2018. The AnnCor CHILDES Treebank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jan Odijk, Margo Zwitserlood-Nijenhuis, Rob Zwitserlood, Jelte van Boheemen, Anouk Scheffer, and Frank Wijnen. 2025. Semi-automatische spontane taalanalyse (SASTA): Op weg naar een efficiënte methode voor de bepaling van het taalprofiel van kinderen met TOS. *Stem-, Spraak- en Taalpathologie*, 30:79–104.

N. Oostdijk, W. Goedertier, F. Van Eynde, L. Boves, J.P. Martens, M. Moortgat, and H. Baayen. 2002. Experiences from the Spoken Dutch Corpus project. In M. González Rodriguez and C. Paz Suárez Araujo, editors, *Proceedings of the third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 340–347. ELRA, Las Palmas.

Benoît Sagot and Éric de la Clergerie. 2006. Error mining in parsing results. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting*

of the Association for Computational Linguistics, pages 329–336, Sydney, Australia. Association for Computational Linguistics.

Liesbeth Schlichting. 2005. *TARSP: Taal Analyse Remediëring en Screening Procedure. Taalontwikkelingsschaal van Nederlandse kinderen van 1-4 jaar*, 7th edition. Pearson, Amsterdam.

Liesbeth Schlichting. 2017. *TARSP: Taalontwikkelingsschaal van Nederlandse kinderen van 1-4 jaar met aanvullende structuren tot 6 jaar*, 8th edition. Pearson, Amsterdam.

Kanta Suzuki, Yoshihide Kato, and Shigeki Matsubara. 2016. Correcting errors in a treebank based on tree mining. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1540–1545, Portorož, Slovenia. European Language Resources Association (ELRA).

Hans van Halteren. 2000. The detection of inconsistency in manually tagged text. In *Proceedings of the COLING-2000 Workshop on Linguistically Interpreted Corpora*, pages 48–55, Centre Universitaire, Luxembourg. International Committee on Computational Linguistics.

Margreet van Ierland, Jeannette Verbeek, and Leen van den Dungen. 2008. *Spontane Taal Analyse Protocol. Handleiding van het STAP-instrument*. UvA, Amsterdam.

Gertjan van Noord. 2004. Error mining for wide-coverage grammar engineering. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 446–453, Barcelona, Spain.

Gertjan van Noord. 2006. At last parsing is now operational. In *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Conférences invitées*, pages 20–42, Leuven, Belgique. ATALA.

Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 147–164. Springer Berlin Heidelberg.

Jeannette Verbeek, Leen van den Dungen, and Anne Baker. 2007. *Spontane Taal Analyse Protocol. Verantwoording van het STAP-instrument, ontwikkeld door Margreet van Ierland*. UvA.

Alexander Volokh and Günter Neumann. 2011. Automatic detection and correction of errors in dependency treebanks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (short papers)*, pages 346–350, Portland, Oregon, USA. Association for Computational Linguistics.

Differentially Private De-identification of Dutch Clinical Notes: A Comparative Evaluation

Michele Miranda[12*], **Xinlan Yan**[34*], **Nishant Mishra**[34], **Rachel Murphy**[34], **Ameen Abu-Hanna**[34], **Sébastien Bratières**[2], **Iacer Calixto**[34]
[1]Sapienza University of Rome, [2]Translated, [3]Amsterdam UMC, [4]University of Amsterdam
miranda@di.uniroma1.it, sebastien@translated.com
{x.yan, n.mishra, r.m.murphy, a.abu-hanna, i.coimbra}@amsterdamumc.nl

Abstract

Protecting patient privacy in clinical narratives is essential for enabling secondary use of healthcare data under regulations such as GDPR and HIPAA. While manual de-identification remains the gold standard, it is costly and slow, motivating the need for automated methods that combine privacy guarantees with high utility. Historically, most automated text de-identification pipelines employed named entity recognition (NER) to identify protected entities for redaction. Although methods based on differential privacy (DP) provide formal privacy guarantees, more recently also large language models (LLMs) are increasingly used for text de-identification in the clinical domain. In this work, we present the first comparative study of DP, NER, and LLMs for *Dutch* clinical text de-identification. We investigate these methods separately as well as hybrid strategies that apply NER or LLM preprocessing prior to DP, and assess performance in terms of privacy leakage and extrinsic evaluation (entity and relation classification). We show that DP mechanisms alone degrade utility substantially, but combining them with linguistic preprocessing, especially LLM-based redaction, significantly improves the privacy–utility trade-off.

Keywords: Clinical Notes, De-Identification, Differential Privacy, NER

1. Introduction

Ensuring privacy in clinical texts is critical to enable data sharing for healthcare research (Conduah et al., 2025). Privacy regulations like GDPR (European Parliament and Council of the European Union, 2016) and HIPAA (U.S. Department of Health and Human Services) require the redaction or *de-identification* of all personally identifiable information (PII) to protect patient privacy.

Methods for PII de-identification based on named entity recognition (NER) have been extensively used for English and other languages (Grouin et al., 2015; Dernoncourt et al., 2017; Bourdois et al., 2021; Tchouka et al., 2022b; Wang et al., 2022). More recently, large language models (LLMs) have been applied to PII de-identification and have shown strong performance (Liu et al., 2023). NER- and LLM-based methods, despite their performance, do not provide any formal privacy guarantees. Differential privacy (DP; Dwork, 2006; Dwork and Roth, 2014), on the other hand, offers a principled mechanism with formal guarantees against privacy leakage when sharing a *privatized* dataset (Chatzikokolakis et al., 2013; Tong et al., 2025).

The tension in the existing literature lies in the fact that DP-based methods for text de-identification provide formal privacy guarantees but can severely impact utility (Yu et al., 2022; Yue et al., 2023; Tchouka et al., 2022b,a), whereas LLM-based methods become increasingly strong but do not provide any privacy guarantees (Pissarra et al., 2024; Yang et al., 2025). These LLM-based methods remain limited as they redact only detected entities rather than full text—offering weaker privacy preservation than DP—, overlook instruction-tuned LLMs, demand computational resources often unavailable in privacy-critical settings, and are developed solely for English and are thus untested in other languages.

In this study, we compare different PII de-

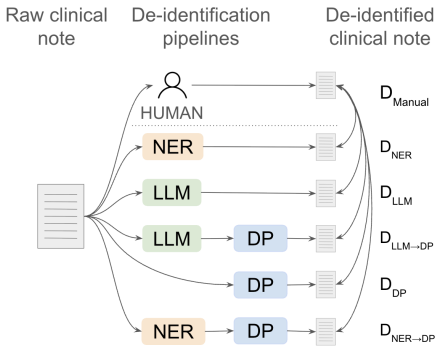


Figure 1: Overview of our comparative analysis. A raw document $\mathcal{D}_{\text{raw}}$ is de-identified using 5 different pipelines, which are evaluated against a manually de-identified version of the same document $\mathcal{D}_{\text{manual}}$. We use a range of open-source and proprietary LLMs that vary in architecture and size in our experiments.

*These authors contributed equally to this work.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 113–124
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

identification methods using state-of-the-art NER (Zaratiana et al., 2024), LLMs (Liu et al., 2023), and synthetic data generation methods with DP guarantees (Chatzikokolakis et al., 2013; Tong et al., 2025) for Dutch clinical notes. See Figure 1 for an overview of our comparisions. Unlike prior works on English or, more recently, French clinical text (Tchouka et al., 2022d,c), our work is **the first study to investigate DP-based text anonymization in Dutch real-world hospital clinical notes**, addressing a critical gap in the de-identification research.

We evaluate the quality of the de-identified clinical notes *intrinsically* by quantifying the remaining residual PII after de-identification, and *extrinsically* by measuring the performance of prediction models trained using the generated de-identified notes for two downstream tasks: entity classification (drugs and disorders) and relation classification (adverse drug events). Finally, we also combine a strong de-identification method (e.g. NER- or LLM-based) as a preprocessing step with DP de-identification. We hypothesise that by doing this we can considerably reduce the DP noise necessary to achieve the same privacy guarantee, possibly leading to a better privacy-utility trade-off.

2. Related Works

De-identification of clinical text is essential to allow secondary use of healthcare data while protecting patient privacy under privacy regulations such as HIPAA (U.S. Department of Health and Human Services) and GDPR (European Parliament and Council of the European Union, 2016). Traditional manual de-identification methods are labour-intensive, costly, and prone to inconsistencies. This section synthesizes key methodological advancements in automated PII de-identification.

Differential Privacy Approaches Differential Privacy (Dwork, 2006) offers formal privacy guarantees by ensuring that the outputs of algorithms are statistically indistinguishable across neighboring datasets. In other words, the key idea is that two models ($\mathcal{M}_1$, $\mathcal{M}_2$) trained on two datasets ($\mathcal{D}_1$, $\mathcal{D}_2$) that differ only by one entry, i.e., one patient record, should be indistinguishable in terms of their predicted outcomes.

Metric DP extends the traditional DP framework by introducing a metric space that quantifies the similarity between data points (Chatzikokolakis et al., 2013). In the context of text data, Metric DP leverages semantic similarity measures to guide the privacy mechanism. Feyisetan et al. (2019) applied Metric DP to word embeddings, enabling controlled perturbations that account for the semantic relationships between words. This results in text transformations that better preserve the utility of the original data while providing privacy guarantees.

Another promising approach is RANTEXT (Tong et al., 2025), a token-level DP framework designed for privacy-preserving text generation. RANTEXT dynamically constructs randomized adjacency lists for context-aware token substitution and has demonstrated strong resistance to membership inference attacks, outperforming earlier models like SANTEXT+ and CUSTEXT+. In empirical evaluations, it achieved up to 98% F1 scores for semantic preservation while maintaining privacy guarantees with ϵ values below 2. Nonetheless, DP-based methods often struggle with high-dimensional clinical narratives, where stricter privacy constraints can significantly degrade data utility (Tong et al., 2025).

LLM-Based De-identification Large language models can be applied to text PII de-identification in different ways. DeID-GPT (Liu et al., 2023), built on GPT-4, asks the model to directly identify and redact sensitive information and integrates HIPAA identifier categories directly into its prompts, achieving 99.25% precision and 89.73% F1 in zero-shot redaction tasks. This can be considered very good performance, especially given the simplicity of the approach.

We note that it is unlikely that organisations would want to use GPT-4 for this task, since using an online model would mean sending sensitive data off-site. Often, this may be even forbidden under privacy regulations. In this work, we use the OpenAI API via a cloud-based service provider in a way that preserves privacy in agreement with privacy regulations. We understand this is not always possible for every hospital, and for that reason we include experiments with both proprietary and open-weight LLMs.

3. Methods and Experiments

Below we detail the dataset (§3.1), the de-identification modules (§3.2) used in our de-identification pipelines that generate privacy-preserving clinical notes (§3.3; Fig. 1), and the intrinsic (§3.4) and extrinsic evaluation on the entity and relation classification tasks (§3.5).

3.1. Dataset

We use the Dutch ADE dataset (Murphy et al., 2025a), a benchmark corpus containing 102 clinical notes from intensive care unit (ICU) patients annotated with entity-level labels—*drug* and *disorder* mentions—and relation labels between *drug-disorder* pairs. The two relation labels annotated are *adverse drug event* or *ADE* (when the drug

Placeholder	Meaning
<AFDELING>	Department
<APOTHEEK>	Pharmacy
<ARTS>	Doctor
<EHR>	Electronic Health Record System
<FEESTDAG>	Holiday
<GEBOORTEDATUM>	Date of Birth
<NAAM>	Name
<RARE_DISEASE>	Rare Disease
<RARE_DISEASE_TREATMENT>	Rare Disease Treatment
<REVALIDATIECENTRUM>	Rehabilitation Center
<SEIN>	Signal
<STAD>	City
<TELNR>	Telephone Number
<TRIAL-ID>	ID of Clinical Trial
<ZIEKENBOEG>	Sickbay
<ZIEKENHUIS>	Hospital
<ZKH>	Abbreviation for Hospital

Table 1: List of placeholders and their meanings.

caused the disorder as an adverse event) and *prescribing indication* (when the drug was prescribed to treat the disorder). We refer to the raw corpus as $\mathcal{D}_{\text{raw}}$ and to its manually de-identified version as $\mathcal{D}_{\text{manual}}$. $\mathcal{D}_{\text{raw}}$ consists of 107,110 words, 0.83% of which are identified as PII. The PII has been redacted and replaced by 17 different placeholders, each indicating one type of PII in the data. There are thus 17 types of PII annotated in $\mathcal{D}_{\text{manual}}$. Names of the placeholders and their meaning are listed in Table 1. For further details on the corpus, we refer the reader to Murphy et al. (2025a).

3.2. De-identification Modules

We apply NER-based, LLM-based, and two DP-based text methods to de-ideintify PII: metric differential privacy (Metric-DP; Chatzikokolakis et al., 2013), and RANTEXT (Tong et al., 2025).

NER-based de-identification We use the off-the-shelf pretrained multilingual NER model `gliner-multi-v2.1`[1], without any fine-tuning on our clinical data. We provide it with our target entity types (see Table 1, e.g., `<NAAM>`, `<ZIEKENHUIS>`) and apply it to extract labeled spans from clinical text. Each prediction consists of text span, label, and confidence score. To balance precision and recall, we use a confidence threshold that maximises the F-1 score on a validation set, i.e., for threshold $t \in [0, 1)$ we only retain entities predicted with confidence score $c \geq t$.

LLM-based de-identification Here, we adapt the approach of Liu et al. (2023) and prompt an LLM in a zero-shot setting to redact predefined PII categories from Table 1 (e.g., `<NAAM>`, `<ZIEKENHUIS>`). When applying LLMs directly as a de-identification pipeline (P_{LLM}), we experiment with several LLMs that vary in architecture, domain specialization, and size, including GPT-4o (Achiam et al., 2023), DeepSeekR1 (8B and 70B) (Guo et al., 2025), LLaMA-3.1 8B (Dubey et al., 2024), and MedGEMMA 27B (Sellergren et al., 2025). Please see Appendix A.1 for the prompt we use. When using LLMs combined with DP ($\text{P}_{\text{LLM}\rightarrow\text{DP}}$), we use the best-performing open-source LLM in terms of privacy (Deepseek-70B).

For our experiments with GPT models, we use the OpenAI API in a privacy-compliant setting whereby no private data leaves the premises of our hospital. Moreover, all our experiments are conducted in a privacy-compliant setting and strictly follow the data usage agreement of the Dutch ADE dataset.

Metric-DP Metric-DP introduces geometric noise in the embedding space to privatize tokens (Chatzikokolakis et al., 2013). We use embeddings from BERTje[2] (De Vries et al., 2019) and construct an approximate nearest-neighbor (ANN) index over the vocabulary.[3] For each token, we sample a noise vector from a γ distribution with scale $1/\epsilon$, add it to the token's embedding, and select the closest replacement from the ANN index.

RANTEXT RANTEXT (Tong et al., 2025) employs an LLM to generate embedding vectors for tokens (we use the Dutch GPT-2[4] from de Vries and Nissim 2020). For each token, a Laplace distribution is used to dynamically determine the size of a random adjacency list, and then new tokens are sampled from this list to replace the original token. Candidate replacements within the noise radius are selected via the exponential mechanism, favoring tokens closer to the original in embedding space.

3.3. De-identification pipelines

We investigate five different *pipelines*, i.e., automatic methods to de-identify the text in the raw dataset ($\mathcal{D}_{\text{raw}}$). Our dataset, as detailed in §3.1, is very small; thus, all pipelines we propose next require models without fine-tuning. The pipelines are: $\mathbf{P}_{\texttt{NER}}$: Apply NER de-identification to $\mathcal{D}_{\text{raw}}$. $\mathbf{P}_{\texttt{LLM}}$: Apply LLM de-identification to $\mathcal{D}_{\text{raw}}$. $\mathbf{P}_{\texttt{DP}}$: Apply DP-based de-identification to $\mathcal{D}_{\text{raw}}$. $\mathbf{P}_{\texttt{NER}\rightarrow\texttt{DP}}$: Apply a NER model to $\mathcal{D}_{\text{raw}}$ and apply DP-based de-

[1] https://huggingface.co/urchade/gliner_multi-v2.1

[2] https://huggingface.co/GroNLP/bert-base-dutch-cased

[3] An ANN index is a data structure designed to quickly retrieve the nearest vectors to a query vector without computing all pairwise distances, which would be prohibitively expensive in high-dimensional spaces.

[4] https://huggingface.co/GroNLP/gpt2-small-dutch

identification to its output. $\mathbf{P}_{\texttt{LLM}\rightarrow\texttt{DP}}$: Apply LLM de-identification to $\mathcal{D}_{\text{raw}}$ and apply DP-based de-identification to its output.

In all experiments with pipeline $\mathbf{P}_{\texttt{LLM}\rightarrow\texttt{DP}}$, we use Deepseek-70B as it is the best-performing open-source LLM in terms of privacy.

3.4. Privacy Leakage Evaluation

We quantify the privacy leakage of a de-identification pipeline as *the percentage of personally identifiable information (PII) leaked in the de-identified text compared to* $\mathcal{D}_{\textit{manual}}$. PIIs are not all the same, and the leakage of different types of PII can, in practice, have very different impact in terms of privacy preservation. For that reason, in this work we differentiate between leakage of *direct* and *indirect PII*.

Direct vs. Indirect PII Direct PII (e.g., a patient's name or telephone number) can directly identify an individual. For that reason, their leakage is the most damaging in terms of privacy preservation. Indirect PII have lower risk but can also be used to identify an individual, usually in combination with other indirect PII or with extra auxiliary information, e.g., Sweeney (2000)'s classical example whereby zip code, date of birth and gender is reported to uniquely identify 87% of the population in the USA.

According to U.S. Department of Health and Human Services, the following identifier types are classified as direct PII: name; address (all geographic subdivisions smaller than state, including street address, city county, and zip code); all date-related elements (except years) related to an individual (including birth date, admission date, discharge date, date of death, and exact age if over 89); telephone number; fax number; email address; social security number; medical record number; health plan beneficiary number; account number; certificate or license number; vehicle identifiers and serial numbers, including license plate numbers; device identifiers and serial numbers; web URL; internet protocol (IP) address; finger or voice print; photographic image; and any other characteristic that could uniquely identify the individual.

In our data, we consider name of doctor, electronic health record ID, date of birth, name of patient, city, and telephone number as direct PIIs, and all the rest in Table 1 as indirect PIIs.

In all our experiments, we apply a strict evaluation setting where the leakage of any subword token part of a PII entity span is treated as leakage.

3.5. Utility Evaluation

We assess the preservation of relevant clinical information by using the generated notes in two downstream tasks: classifying entity spans in a clinical text as drug or disorder mentions, and predicting whether a drug–disorder entity pair denotes an adverse drug event (ADE).[5] We use the original train/dev/test splits in the Dutch ADE corpus for training, model selection, and testing, respectively. For configuration details please see Appendix A.2 We use the annotated entities and relations that remain in the generated texts for training and development, and evaluate the model against the original test set with gold annotations. At low ϵ values under some settings, i.e. $\epsilon\leq$128 for metric-DP and $\epsilon\leq$16 for RANTEXT, we only have less than 10 annotated relations in the generated text, making it impossible to train the relation classification model. Thus, we omit these experiments and report performance using macro F-1 score for both entity and relation classification. We do not report new confidence intervals for the downstream tasks because its protocol and uncertainty were established in prior work (Murphy et al., 2025b), which already provides CIs under the same evaluation procedure.

4. Results

We now report on our experiments regarding the empirical privacy leakage of PII in the generated clinical notes (4.1), and the downstream evaluations according to drug and disorder (entity) classification (4.2) and adverse drug event (relation) classification (4.3).

4.1. Privacy Leakage Results

Figure 2 presents privacy leakage percentage for various pipeline configurations across DP budgets (ϵ), spanning from 8 to 1024. We acknowledge that with ϵ higher than 10, the theoretical guarantees of DP fade away, but it is also usual in practical setting to work with high epsilons in order to get meaningful empirical results, as seen in works like DP-BART (Igamberdiev and Habernal, 2023). As ϵ increases, privacy leakage also rises, demonstrating the trade-off between reduced noise (higher ϵ) and weaker privacy preservation. At high privacy budgets, the protection provided by DP diminishes, leading to increased privacy leakage across all pipelines. Among the DP-based pipelines, $\mathrm{P}_{\texttt{DP}}$ consistently exhibits the highest privacy leakage across all ϵ values. In contrast, pipelines that combine DP with NER preserve privacy for slightly higher privacy budgets, and LLM-based de-identification show significantly lower empirical leakage. Notably,

[5] We note that originally the Dutch ADE dataset was collected to support these two tasks (Murphy et al., 2025b), making these tasks ideal for utility evaluation.

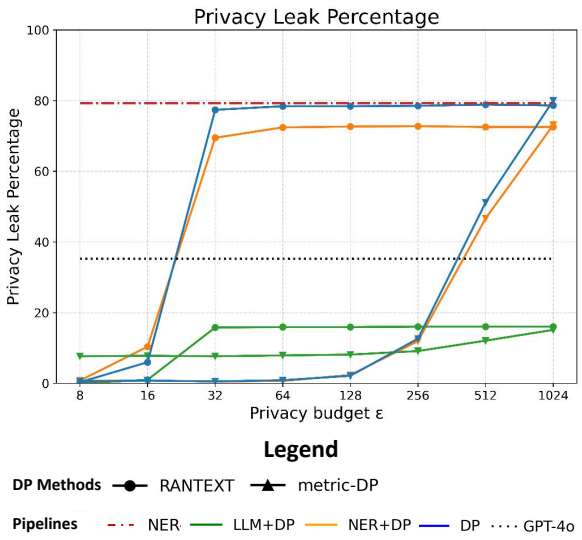


Figure 2: Comparison of privacy leakage across different de-identification pipelines and DP budgets (ϵ). This figure includes two DP mechanisms: RANTEXT and Metric-DP, each applied to three pipelines: P_{DP}, $P_{NER \to DP}$, and $P_{LLM \to DP}$. For $P_{LLM \to DP}$, we use Deepseek-70B as the de-identification module as it performs the best in terms of privacy. Horizontal lines indicate non-DP baselines, including one NER-based pipeline (GLiNER) and one best performing LLM (GPT-4o). As ϵ increases, privacy leakage of applying DP directly to raw text also significantly increases. Meanwhile, combining DP with LLM-based redaction largely enhances privacy even with high ϵ values.

pipelines with LLM preprocessing significantly outperform other approaches with a high budget. This improvement can be attributed to the strong ability of LLMs to identify and redact sensitive information more effectively, reducing the amount of private data exposed to DP noise.

In our next experiments, we use the Health Insurance Portability and Accountability Act (HIPAA; Garfinkel et al., 2015) to differentiate between leakage of *direct* (i.e., exact identifiers such as names, emails, phone numbers) versus *indirect PII* (i.e., *quasi-identifiers* and other attributes that enable re-identification when linked). Figure 3 illustrates the leakage rates of direct and indirect PIIs for each privacy budget ϵ across pipelines. Overall, leakage increases with ϵ, but the extent and composition differ largely according to the pipeline. Metric-DP consistently leaks far less PII than RANTEXT at comparable budgets: while RANTEXT and NER+RANTEXT reach 38–40% direct (and 76–71% total) leakage at high ϵ, metric-DP (MDP) better preserves privacy more at low-to–mid ϵ, and only approaches RANTEX when ϵ is large. Moreover, leakage rises with ϵ for all methods, but the slope differs. RANTEXT variants escalate quickly, with the direct component dominating the increase, whereas MDP grows more gradually until the largest budgets, where its bars steepen.

Moreover, the baseline results of LLMs and NER confirms that stand-alone methods without DP are markedly leakier: the pure NER system shows high overall leakage with a large direct component, and LLM-only variants also leak substantially. However, inserting an LLM de-identification before the DP step shows a clear advantage for both families. Finally, pipelines combining LLMs with DP (LLM+MDP and LLM+RANTEXT) were the only pipelines we investigated that managed to keep empirical privacy leakage below 10%, the majority of which consisted of indirect PII.

4.2. Utility Preservation Results - Entity Classification

Figure 4 illustrates the relationship between privacy budgets (ϵ) and entity classification F1 scores (utility) across different pipelines. At very low privacy budgets ($\epsilon \leq 32$), all DP pipelines demonstrate poor utility due to the high levels of noise, but utility improves as ϵ increases, peaking around $\epsilon = 512$. P_{DP} consistently exhibits lower F1 scores across all privacy budgets, demonstrating utility degradation caused by noise addition, particularly at low ϵ values. Introducing linguistic preprocessing improves utility, with $P_{NER \to DP}$ achieving higher F1 scores than P_{DP}, especially at mid-to-high privacy budgets ($\epsilon \geq 128$). $P_{LLM \to DP}$ further outperforms $P_{NER \to DP}$ in most cases, showcasing the advantage of leveraging advanced language models for preprocessing. Both DP methods (RANTEXT and metric-DP) exhibit similar trends within each pipeline, suggesting that preprocessing, rather than the DP mechanism itself, plays a more critical role in improving utility. The LLM baseline maintains consistently high F1 scores near 1.0, thereby serving as an upper bound for utility.

4.3. Utility Preservation Results - Relation Classification

Figure 5 shows relation classification F1 scores (utility) across various privacy budgets (ϵ) and for different pipelines. At low privacy budgets, such as $\epsilon = 8$ or in some cases 16, no results are obtained due to the severe obfuscation which reduces the amount of usable training data to negligible levels. As ϵ increases, utility scores gradually improve across all pipelines. P_{DP} exhibit the lowest F1 scores throughout the range of privacy budgets. In contrast, both $P_{NER \to DP}$ and $P_{LLM \to DP}$ achieve higher F1 scores, with $P_{LLM \to DP}$ consistently outperforming $P_{NER \to DP}$, particularly at mid-to-high privacy budgets ($\epsilon \geq 128$). Non-DP baselines still maintain

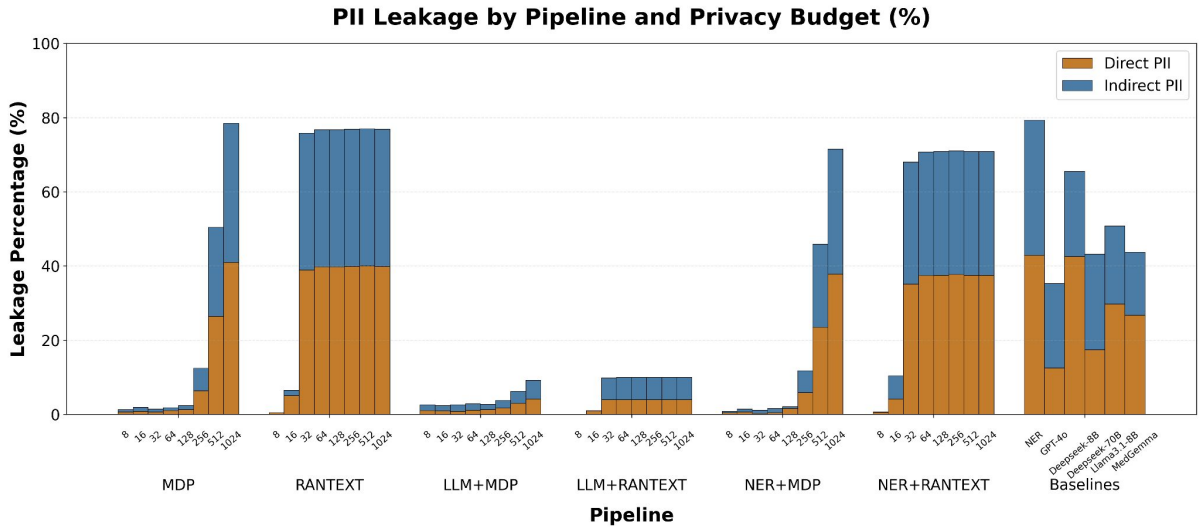


Figure 3: PII leakage by pipeline and privacy budget ϵ. Bars are stacked (dark: direct PII; light: indirect PII). Across budgets, metric-DP (MDP) leaks substantially less than RANTEXT, and all methods show increasing leakage as ϵ grows, with the steepest rise for RANTEXT variants. Prepending an LLM rewrite reduces leakage for both families, with LLM+MDP achieving the lowest overall rates, followed by LLM+RANTEXT, while plain RANTEXT and NER+RANTEXT exhibit high direct leakage. NER and LLMs alone without DP also remain markedly leakier.

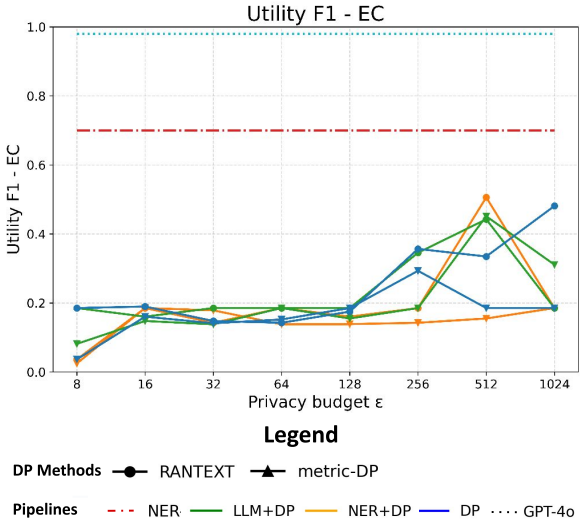


Figure 4: Comparison of utility F1-score for Entity Classification (EC) task across different de-identification pipelines and DP budgets (ϵ) (see Figure 2 for more details). It shows that the utility scores improve with higher ϵ, but still cannot recover baseline performances.

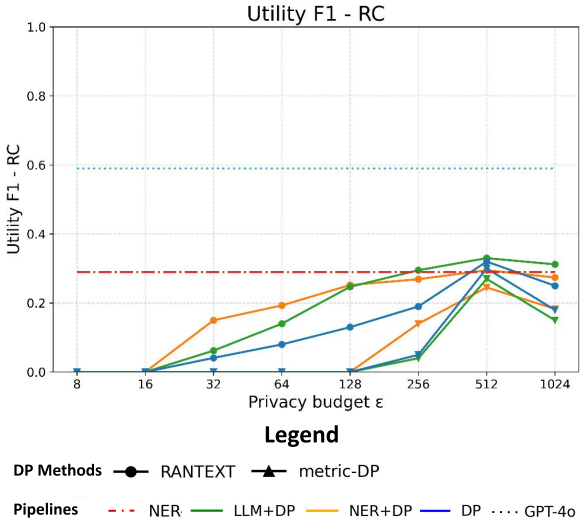


Figure 5: Comparison of utility F1-score for Relation Classification (RC) task across different de-identification pipelines and DP budgets (ϵ) (see Figure 2 for more details). It shows that the utility scores slightly improve with higher ϵ, but are significantly harmed even with high ϵ, and cannot recover baseline performances.

consistently high F1 scores and establish the upper bounds for utility. These results underscore the crucial role of linguistic preprocessing, particularly with LLMs, in mitigating the utility degradation inherent to DP-based methods, especially in tasks requiring complex relational reasoning.

4.4. LLM Baseline Results

Figure 6 shows the performances of privacy leakage and utility F1-scores for both EC and RC tasks among P_{LLM} using 5 various LLMs. Amongst the LLM baselines, GPT-4o achieves the best, i.e. lowest privacy leakage percentage while still offering

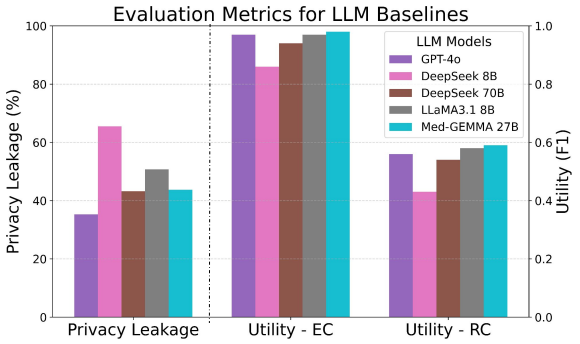


Figure 6: Comparison of evaluation metrics including Privacy Leakage, Utility - EC, and Utility - RC for P_{LLM} using 5 different LLMs. Privacy Leakage, where lower is better, shows that GPT-4o achieves the lowest leakage. Utility metrics for EC and RC, where higher is better, reveal that Med-GEMMA 27B consistently outperforms others while obtaining relatively low privacy leakage. GPT-4o achieves similar utility even with the lowest privacy leakage, indicating its high overall performance.

near-optimal utility for both EC and RC tasks. Med-GEMMA 27B exhibits the highest utility for both tasks while keeping the privacy leakage at a relatively low level. This aligns with the assumption that LLMs' advanced linguistic capabilities allow them to effectively redact sensitive information.

4.5. Operational Efficiency

Building the manually de-identified corpus ($\mathcal{D}_{\text{manual}}$) was a labor-intensive process, requiring over 47 hours of annotation by two annotators (Murphy et al., 2025a). In contrast, our automated pipelines are way more time- and cost-efficient, with metric-DP taking 3 minutes and 16 seconds per run and RANTEXT 26 minutes and 40 seconds per run. While manual de-identification remains the gold standard for accuracy and consistency, these results highlight a substantial operational advantage for the automated methods, enabling rapid iteration over various configurations and large-scale or repeated de-identification in practice.

5. Discussion and Conclusions

General results. This work provides a comparative analysis of privacy-preserving de-identification for Dutch clinical notes by contrasting text-level differential privacy (DP) mechanisms, NER-based masking, and LLM-based masking, and by evaluating both privacy leakage and downstream utility on entity and relation classification tasks. Our results show a consistent tension for DP perturbation: increasing the privacy budget ϵ improves utility but increases residual leakage, and applying DP directly to raw notes yields the weakest privacy outcomes among DP pipelines. By contrast, preprocessing that removes explicit identifiers before DP (via NER or LLM masking) improves the overall trade-off, indicating that, in our no-training setting, most gains come from reducing sensitive content *before* perturbation rather than relying on perturbation alone. In utility terms, token-level perturbation is especially harmful for relation classification (more than for entity classification), plausibly because relations depend on longer-range coherence and on the availability of sufficient high-quality labeled instances: under tighter privacy budgets, the perturbation can effectively corrupt the signal beyond recovery, whereas higher ϵ only partially closes the gap with non-private baselines. Finally, our operational analysis highlights the substantial gap between the cost of producing benchmark-quality manually de-identified reference data and the speed of automated pipelines, which enables rapid iteration over configurations in practice.

Advice for practitioners. When the goal is to share Dutch clinical narratives while explicitly controlling privacy risk, our results suggest prioritizing hybrid strategies: use a strong high-recall masking stage (LLM-based when feasible, otherwise NER-based) and then apply DP as a second-stage safeguard on the residual text, rather than applying DP directly to raw notes. This design reduces the sensitive surface exposed to DP and consistently improves both leakage and downstream utility relative to DP-on-raw-text, while retaining an explicit privacy knob through ϵ. Practitioners should also anticipate that preserving utility for relation-centric tasks will be harder than for entity-centric tasks under text perturbation, and should validate performance on the intended downstream use case (not only on intrinsic leakage). Finally, even when automated pipelines are adopted, a manually de-identified reference (or equivalent audit protocol) remains important for trustworthy evaluation and monitoring. Although the best-performing LLM in our experiments is proprietary (GPT-4o), recent work on Italian clinical notes has shown that smaller, open-source LLMs can achieve competitive de-identification performance in a similar zero-shot setting (Miranda et al., 2025), suggesting that our hybrid pipeline design is not inherently tied to proprietary models.

Future work. Three extensions are particularly important: (i) expanding beyond Dutch to assess whether the same trade-offs hold across languages with different resource availability; (ii) enabling broader experimentation and external validation via public benchmarks or carefully designed synthetic alternatives; and (iii) widening utility evaluation be-

yond entity/relation classification to additional clinical NLP use cases (e.g., document-level prediction), where sensitivity to perturbation may differ. More generally, future work should also strengthen adversarial evaluation of re-identification risk and explore hybrid designs that better target quasi-identifiers without unnecessarily corrupting clinical signals.

Limitations

While our study provides valuable insights into the application of differential privacy (DP) for clinical text de-identification, it has certain limitations. First, our work focuses exclusively on Dutch clinical narratives, which may limit the generalizability of our findings to other languages or multilingual clinical datasets. Expanding this research to additional languages would provide a broader understanding of the effectiveness and scalability of the proposed techniques. Second, the dataset used in this study is private, as it contains sensitive medical information, which restricts reproducibility and external validation by other researchers. Although this is a common constraint in clinical research, the use of publicly available or synthetic benchmark datasets in future studies could help foster more extensive experimentation and comparison. Third, our privacy leakage metric - residual PII percentage against a manually de-identified reference - is an empirical proxy, not a formal privacy measure. While DP provides theoretical guarantees on output indistinguishability, empirical leakage captures a complementary quantity: how much identifiable information survives in practice. These two notions should not be conflated, yet empirical evaluation against a gold-standard reference remains one of the few viable approaches for assessing de-identification effectiveness on real data. Lastly, while we evaluated our pipelines on two downstream tasks (entity and relation classification), further exploration of their performance across a wider variety of use cases, such as document classification or clinical outcome prediction, would provide a more comprehensive assessment of their utility.

6. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Loick Bourdois, Marta Avalos, Gabrielle Chenais, Frantz Thiessard, Philippe Revel, Cedric Gil-Jardine, and Emmanuel Lagarde. 2021. De-identification of emergency medical records in french: Survey and comparison of state-of-the-art automated systems. *The International FLAIRS Conference Proceedings*, 34.

Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies*.

Andrew Kweku Conduah, Sebastian Ofoe, and Dorothy Siaw-Marfo. 2025. Data privacy in healthcare: Global challenges and solutions. *Digital Health*, 11:20552076251343959.

Wietse de Vries and Malvina Nissim. 2020. As good as new. how to successfully recycle english gpt-2 to make models for other languages.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association (JAMIA)*, 24(3):596–606.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, page 1–12, Berlin, Heidelberg. Springer-Verlag.

Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407.

European Parliament and Council of the 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 (general data protection regulation). *Official Journal of the European Union*, L119:1–88.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text.

2019 IEEE International Conference on Data Mining (ICDM), pages 210–219.

Simson Garfinkel et al. 2015. De-identification of personal information:.

Cyril Grouin, Nicolas Griffon, and Aurélie Névéol. 2015. Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs? In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 31–39, Lisbon, Portugal. Association for Computational Linguistics.

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

Timour Igamberdiev and Ivan Habernal. 2023. DP-BART for privatized text rewriting under local differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934, Toronto, Canada. Association for Computational Linguistics.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, Fang Zeng, Lichao Sun, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4.

Michele Miranda, Sébastien Bratières, Stefano Patarnello, and Livia Lilli. 2025. Mamma mia! where's my name? de-identifying Italian clinical notes with large language models. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 735–746, Cagliari, Italy. CEUR Workshop Proceedings.

Rachel M Murphy, Dave A Dongelmans, Nicolette F de Keizer, Rosa J Jongeneel, Christiaan H Koster, Kitty J Jager, Ameen Abu-Hanna, Iacer Calixto, and Joanna E Klopotowska. 2025a. Creation of a gold standard dutch corpus of clinical notes for adverse drug event detection: the dutch ade corpus. *Language Resources and Evaluation*, pages 1–17.

Rachel M Murphy, Nishant Mishra, Nicolette F de Keizer, Dave A Dongelmans, Kitty J Jager, Ameen Abu-Hanna, Joanna E Klopotowska, and Iacer Calixto. 2025b. Detection of adverse drug events in dutch clinical free text documents using transformer models: benchmark study. *arXiv preprint arXiv:2507.19396*.

David Pissarra, Isabel Curioso, João Alveira, Duarte Pereira, Bruno Ribeiro, Tomás Souper, Vasco Gomes, André Carreiro, and Vitor Rolla. 2024. Unlocking the potential of large language models for clinical text anonymization: A comparative study. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 74–84, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34.

Yakini Tchouka, Jean-François Couchot, Maxime Coulmeau, David Laiymani, Philippe Selles, and Azzedine Rahmani. 2022a. De-identification of french unstructured clinical notes for machine learning tasks. *arXiv preprint arXiv:2209.09631*.

Yakini Tchouka, Jean-François Couchot, and David Laiymani. 2022b. An easy-to-use and robust approach for the differentially private de-identification of clinical textual documents. *arXiv preprint arXiv:2211.01147*.

Yakini Tchouka, Jean-François Couchot, Maxime Coulmeau, David Laiymani, Philippe Selles, and Azzedine Rahmani. 2022c. De-identification of french unstructured clinical notes for machine learning tasks. *arXiv preprint arXiv:2209.09631*.

Yakini Tchouka, Jean-François Couchot, and David Laiymani. 2022d. An easy-to-use and robust approach for the differentially private de-identification of clinical textual documents. *arXiv preprint arXiv:2211.01147*.

Meng Tong, Kejiang Chen, Jie Zhang, Yuang Qi, Weiming Zhang, Nenghai Yu, Tianwei Zhang, and Zhikun Zhang. 2025. Inferdpt: Privacy-preserving inference for black-box large language model.

U.S. Department of Health and Human Services. 45 cfr § 164.514 – de-identification of health information. Health Information Privacy. [Online]. Available: https://www.law.cornell.edu/cfr/text/45/164.514. [Accessed: Dec. 2, 2024].

Stella Verkijk and Piek Vossen. 2021. Medroberta. nl: a language model for dutch electronic health records. In *Computational Linguistics in the*

Netherlands, volume 11, pages 141–159. Computational Linguistics in the Netherlands.

Peng Wang, Yong Li, Liang Yang, Simin Li, Linfeng Li, Zehan Zhao, Shaopei Long, Fei Wang, Hongqian Wang, Ying Li, and Chengliang Wang. 2022. An efficient method for deidentifying protected health information in chinese electronic health records: Algorithm development and validation. *JMIR Med Inform*, 10(8):e38154.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. 2025. Robust utility-preserving text anonymization based on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28922–28941, Vienna, Austria. Association for Computational Linguistics.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2022. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, Toronto, Canada. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

A. Appendix

A.1. LLM De-identification Prompts and Performances

We use the following prompt for using LLMs as de-identification tools: "*Please anonymize the following clinical note. Replace all the following information with corresponding masks: Replace hospital names with <ZIEKENHUIS>; Replace abbreviations for hospitals (e.g. ZKH) with <ZKH>; Replace doctor names with <ARTS>; Replace city names with <STAD>; Replace signs or signals (SEIN) with <SEIN>; Replace patient names with <NAAM>; Replace entries from electronic health records with <EHR>; Replace rare diseases with <RARE_DISEASE>; Replace clinical trial identifiers with <TRIAL-ID>; Replace holidays with <FEESTDAG>; Replace rare disease treatments with <RARE_DISEASE_TREATMENT>; Replace department names with <AFDELING>; Replace rehabilitation centers with <REVALIDATIECENTRUM>; Replace sickbay references with <ZIEKENBOEG>; Replace telephone numbers with <TELNR>; Replace pharmacies or drug stores with <APOTHEEK>; Replace dates of birth with <GEBOORTEDATUM>.*"

Model	F1-score	TP	FP	FN
GPT-4o	0.205839	569	4848	310
DeepSeek-8b	0.337265	303	718	576
DeepSeek-70b	0.337807	499	3614	380
LLaMA3-8b	0.492872	433	652	446
MedGemma	0.591672	495	399	384

Table 2: Comparison of LLMs by F1-scores, True Positives (TP), False Positives (FP), and False Negatives (FN).

Table 2 provides a comparison of macro F1-scores, true positives (TP), false positives (FP), and false negatives (FN) across various LLM models applied to the redaction task. DeepSeek-8b and DeepSeek-70b exhibit relatively low F1-scores of 0.337265 and 0.337807, respectively, with DeepSeek-70b achieving slightly higher true positives (TP = 499) but at the cost of significantly more false positives (FP = 3614). LLaMA3-8b demonstrates a notable improvement with a good balance between true positives and false positives. MedGemma emerges as the best-performing model, with the highest macro F1-score of 0.591672. This is supported by its significant reduction in false positives and the highest true positives.

A.2. Downstream Evaluation Configuration

For the downstream Entity Classification (EC) and Relation Classification (RC) tasks, we use MedRoberta.nl (Verkijk and Vossen, 2021) as the base model due to its best performance according to Murphy et al. (2025b) and train with the remaining entity/relation in our generated data. We use a batch size of 8, a learning rate of 3e-5, a warmup ratio of 0.2 for EC and a batch size 128, hidden sizes of 512, 128, 32, a dropout rate of 0.5, a learning

rate 1e-6, and a patience of 30 for RC. We report the micro F1-score of EC and macro F1-score of RC as annotated relations suffer from high imbalance. Table 3 and Table 4 show the full results of the EC and RC evaluation.

Pipeline/Epsilon	8	16	32	64	128	256	512	1024	∞
Baseline	–	–	–	–	–	–	–	–	0.91
NER	–	–	–	–	–	–	–	–	0.70
LLM	–	–	–	–	–	–	–	–	0.97
metricDP	0.0368	0.1605	0.1412	0.1524	0.1854	0.2933	0.1854	0.1854	–
RANTEXT	0.1854	0.1898	0.1479	0.1427	0.1751	0.3567	0.3348	0.4812	–
NER+metricDP	0.0249	0.1855	0.1787	0.1383	0.1388	0.1427	0.1552	0.1854	–
NER+RANTEXT	0.0368	0.1854	0.1427	0.1854	0.1605	0.1854	0.5061	0.1854	–
LLM+metricDP	0.0817	0.1479	0.1383	0.1854	0.1552	0.1854	0.4514	0.3110	–
LLM+RANTEXT	0.1854	0.1605	0.1854	0.1854	0.1854	0.3457	0.4426	0.1854	–

Table 3: Performance of Entity Classification task across privacy budgets (ϵ). Dashes indicate settings where ϵ is not applicable.

Pipeline/Epsilon	8	16	32	64	128	256	512	1024	∞
Basline	–	–	–	–	–	–	–	–	0.62
NER	–	–	–	–	–	–	–	–	0.29
LLM	–	–	–	–	–	–	–	–	0.59
metricDP	–	–	–	–	–	0.050	0.300	0.180	–
RANTEXT	–	–	0.041	0.080	0.130	0.190	0.320	0.250	–
NER+metricDP	–	–	–	–	–	0.140	0.245	0.183	–
NER+RANTEXT	–	–	0.150	0.193	0.252	0.269	0.295	0.274	–
LLM+metricDP	–	–	–	–	–	0.040	0.270	0.150	–
LLM+RANTEXT	–	–	0.062	0.140	0.247	0.295	0.330	0.312	–

Table 4: Performance of Relation Classification task across privacy budgets (ϵ). Dashes indicate settings where ϵ is not applicable.

SQUCS at MEDIQA-SYNUR 2026: A Multi-Agent Open Source LLM System for Nursing Observation Extraction

Riham Jeeballah[1], **Adhari AlZaabi**[2], **and Abdulrahman K. AAlAbdulsalam**[1]
[1]Sultan Qaboos University, Department of Computer Science, Oman
[2]Sultan Qaboos University, Department of Clinical and Human Anatomy, Oman
a.aalabdulsalam@squ.edu.om

Abstract

Clinical nursing documentation contains detailed observational information that is essential for patient monitoring and clinical decision-making, yet this information is predominantly recorded in free-text form. The MEDIQA-SYNUR shared task addresses this challenge by requiring systems to extract structured nursing observations from clinical transcripts under strict constraints on evidence grounding and value normalization. In this work, we present a multi-agent large language model (LLM)–based system for the MEDIQA-SYNUR task. We utilize the Llama3 open source LLM for this purpose for ease of local deployment within hospital digital infrastructure. Our system decomposes the extraction process into specialized agents responsible for schema-guided extraction, rule-based validation, and precision-oriented filtering. Starting from a baseline multi-agent pipeline, we conduct a systematic error analysis over the entire development set, examining all false positive and false negative predictions. Our final configuration, selected after extensive exploration and error analysis, combined transcript segmentation, the precision agent, and a suppression table derived from development-set analysis. On the development set, this setup achieved an F1 score of 0.6930 (precision = 0.6427, recall = 0.7518). Applying the same configuration directly to the test set, without any additional tuning, yielded an F1 score of 0.5923 (precision = 0.5292, recall = 0.6725). These results represent the most effective balance of precision and recall achieved through our iterative refinements and reflect the final state of the system as submitted for the competition.

Keywords: Clinical Natural Language Processing, Nursing Observation Extraction, Multi-Agent Systems, Large Language Models, MEDIQA-SYNUR, Evidence-Grounded Extraction

1. Introduction

Extracting information from clinical text reports has been an active area of research within the last two decades (Wang et al., 2018; Meystre et al., 2008). Converting unstructured narrative text into structured schema is important for medical research and effective clinical decision making (Kreimeyer et al., 2017).

The majority of studies focused on extracting information from clinical documentation produced by physicians, and recently more publications investigated extracting information specifically from nursing notes (Mitha et al., 2023). Most recent studies focused on identifying one particular information from text authored by nurses such as ICU length of stay and mortality from intensive care units (Huang et al., 2021), risk of fall (Bjarnadottir and Lucero, 2018), sexual orientation and gender, and presence of urinary catheter and related symptoms (Gundlapalli et al., 2017).

The SYNUR (Synthetic NURsing) dataset, on the other hand, was proposed to cover a variety of comprehensive clinical observations often found within nursing dictations (Corbeil et al., 2025a). In this study, we describe an approach that extracts structured observations from the SYNUR dataset as part of the MEDIQA-SYNUR shared task (George Michalopoulos, 2026). The best submission achieved an F1 score of 0.59 with recall of 0.67 and precision of 0.53 on the test set. In this study we describe the development of multi-agent LLM-based system for this task.

2. Related Work

Structured extraction from clinical free text has been approached through rule-based systems, machine learning, and, more recently, transformer-based models. A systematic review of 127 papers covering NLP methods applied to EHRs found that named entity recognition and clinical note classification now dominate the application landscape, with deep learning and transfer learning architectures increasingly applied across these tasks (Hossain et al., 2023). While these advances have improved extraction from physician-authored documents such as radiology reports and discharge summaries, nursing documentation poses distinct challenges that limit the direct transfer of existing methods. Unlike physician notes, which favour concise problem-oriented abstraction, nursing narratives are illustrative and descriptive, rely on institution-specific phrasing, and document a broader range of observations per encounter (Hyun et al., 2009). An integrative review of 43 NLP studies using nursing notes further confirmed that methodological approaches remain heterogeneous and that nursing-specific standard terminolo-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 125–135
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

gies are used in fewer than one in five published studies (Mitha et al., 2023). A scoping review of NLP applied in post-acute care nursing settings found that existing systems target a narrow range of risk outcomes, and many studies do not employ standardized terminologies (Scharp et al., 2024). Taken together, these gaps point to the absence of comprehensive, schema-grounded benchmarks for nursing observation extraction prior to MEDIQA-SYNUR dataset.

The emergence of large language models (LLMs) has introduced a new paradigm for clinical information extraction, enabling schema-guided structured output without task-specific fine-tuning. Multi-step LLM pipelines that combine an LLM extraction stage with rule-based post-processing have shown particular promise. (Wang et al., 2024) applied such an architecture to medical record entity extraction, using LLM-generated question templates followed by rule-based filtering to suppress extraction inaccuracies, demonstrating that decomposing the extraction task across specialised stages improves schema alignment and reduces spurious predictions. (Adam et al., 2025) further showed that targeted prompting strategies combined with post-hoc validation heuristics can substantially suppress hallucinated numeric extractions from clinical text, a finding that directly informs the design of our validation and precision filtering agent. A critical practical constraint for hospital deployment is data privacy: proprietary cloud-hosted LLMs require transmitting protected health information to external servers, raising significant legal and ethical barriers. (Wiest et al., 2024) demonstrated that a locally deployed Llama 2 pipeline achieves strong zero-shot clinical extraction performance, with sensitivity and specificity exceeding 90% for key features, while keeping all patient data on-premise. This motivates our selection of LLaMA-3.3 for local deployment within hospital infrastructure.

Within the MEDIQA shared task series, which has benchmarked clinical NLP systems annually since 2019 across tasks including question answering, dialogue-to-note summarization, and medical order extraction (Abacha et al., 2019, 2021; Bjarnadottir et al.; Abacha et al., 2023; Saeed, 2024; Corbeil et al., 2025b), top-performing systems have progressively shifted from fine-tuned encoder models toward generative LLM-based pipelines. The MEDIQA-SYNUR 2026 task is distinctive within this series in requiring verbatim evidence grounding and value normalization across 192 nursing-specific schema categories constraints that no prior shared task has combined (George Michalopoulos, 2026).

3. Materials and Methods

3.1. Dataset

We use the MEDIQA-SYNUR dataset released on Codabench, which consists of clinical nursing transcripts paired with structured annotations. The dataset includes 122 training records, 101 development records, and 199 test records.

The transcripts follow a conversational clinical style, introducing ambiguity and variability that make structured extraction challenging. All annotations are strictly grounded in the transcript text, and systems are required to extract only explicitly stated observations with verbatim evidence spans.

Additional dataset statistics and distribution analyses are provided in the Appendix.

3.2. Observation Schema

The MEDIQA-SYNUR task defines a comprehensive observation schema consisting of 192 distinct nursing observation categories, each identified by a unique concept ID. These categories are designed to capture a wide spectrum of nursing assessments, interventions, physiological measurements, and patient status descriptors commonly documented in clinical care settings. Each observation category is associated with a predefined value type, which constrains the allowable output format and guides system extraction. The schema includes four primary value types: `SINGLE_SELECT` (e.g., Skin turgor: {Tented, Normal}); `MULTI_SELECT` (e.g., Urinary symptoms: {urgency, difficulty urinating}); `NUMERIC` (e.g., Urine output, Oxygen saturation); and `STRING` (e.g., Pain description, Heart sounds).

The schema spans multiple clinical domains, including but not limited to cognitive and neurological status (e.g., Cognitive status, Orientation, Glasgow Coma Scale components), respiratory and cardiovascular assessments (e.g., Oxygen delivery device, Work of breathing, Heart rate), mobility and functional status (e.g., Mobility, Level of assistance), genitourinary and gastrointestinal observations (e.g., Urinary symptoms, Urine appearance, Bowel movement description), skin and wound assessments (e.g., Skin turgor, Edema, Pressure injury stage), as well as nursing interventions and safety-related checks (e.g., Intravenous therapy, Bed safety, Fall risk assessment).

In addition, the schema explicitly models measurement units as separate observation categories for several numeric concepts (e.g., Urine output unit, Oxygen saturation unit, Temperature unit). This design requires systems to independently extract both the numeric value and its corresponding unit when present in the transcript, increasing task granularity and precision.

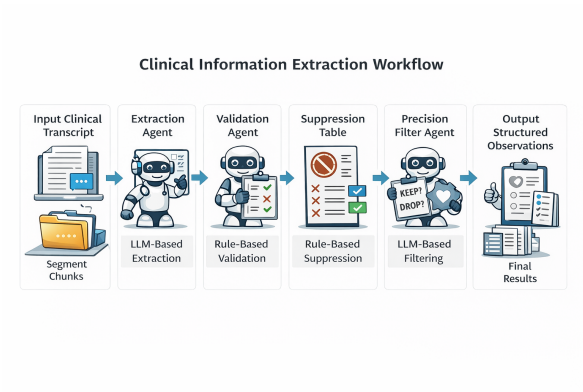


Figure 1: Overview of the proposed multi-agent clinical information extraction pipeline. The input clinical transcript is first segmented into smaller chunks for scalable processing. Each segment is processed by an LLM-based extraction agent to generate candidate observations, followed by a rule-based validation agent. A deterministic suppression table, derived from error analysis, is applied to reduce systematic false positives. A final LLM-based precision filter performs conservative KEEP/DROP decisions before aggregating results into structured observations.

3.3. Multi-Agent System Architecture

The proposed system follows a multi-stage, multi-agent architecture designed to extract structured nursing observations from free-text clinical transcripts while strictly enforcing evidence-based constraints. Each agent addresses a specific failure mode observed during development, including hallucinated extractions, implicit inference, and systematic false positives. The pipeline consists of four main components: an extraction agent, a validation agent, a precision filtering agent, and a deterministic suppression mechanism.

3.3.1. Overview of the Pipeline

Figure 1 illustrates the overall workflow of the proposed multi-agent system.

The system follows a sequential multi-stage design in which each component addresses a specific failure mode observed during development. Starting from transcript segmentation, each chunk is processed independently by the extraction agent, followed by validation, suppression, and precision filtering stages. The final outputs from all chunks are aggregated into structured observations.

Transcript Segmentation

Clinical transcripts in MEDIQA-SYNUR may be lengthy and information-dense. To reduce context overload and improve coverage, transcripts are optionally segmented into smaller text chunks based on character length (Corbeil et al., 2025c). Each segment is processed independently, allowing the extraction agent to focus on localized context and improving recall for observations mentioned sparsely or late in the transcript.

Schema Segmentation

A. Schema Segmentation and Batching (Static Schema Partitioning)

The MEDIQA-SYNUR observation schema contains 192 distinct categories, making it impractical to present the full schema to the model in a single prompt. To handle this, we implement a static batching strategy, where the schema categories are divided into fixed batches. Each batch is processed independently by the extraction agent, and only the subset of schema definitions belonging to that batch is included in the prompt.

B. Schema Retrieval via Embeddings (Dynamic Top-K Selection)

Each schema category is embedded once using a text-embedding model. At inference time, the transcript is also embedded, and cosine similarity is used to retrieve the top-K most relevant schema categories. Only these retrieved definitions are injected into the prompt for extraction.

This dynamic retrieval mechanism reduces prompt length, focuses the model on clinically relevant categories, and adapts naturally to transcript variability. However, experimental results showed that this retrieval-based strategy provided limited performance gains compared to full schema batching. Therefore, it was not included in the final system configuration.

3.3.2. Extractor Agent

The Extractor Agent is responsible for the initial identification of candidate nursing observations from the transcript using a large language model (LLM). To manage the large schema size (192 observation categories), the schema is dynamically chunked into smaller batches, and each batch is processed independently against segmented transcript chunks when segmentation is enabled. The extraction prompt explicitly enforces strict evidence rules. Each extracted observation must include a valid schema identifier, a value consistent with the schema type (numeric, single-select, multi-select, or free text), and an evidence field that is an exact verbatim substring copied from the transcript. The prompt explicitly prohibits inference, speculation, and abstraction. Negative values (e.g., “No”, “Absent”) are only allowed when explicit negation cues appear in the transcript. The output of the extraction agent is intentionally permissive, prioritizing recall while deferring strict correctness checks to downstream agents.

3.3.3. Validation Agent

The Validator Agent applies deterministic, rule-based validation to the raw outputs produced by the extraction agent. Its primary role is to remove structurally invalid, weakly grounded, or speculative observations before further refinement. Validation rules include rejecting observations whose evidence does not appear verbatim in the transcript, filtering out evidence containing hedging or speculative language (e.g., “possibly”, “suggests”, “likely”), enforcing explicit negation rules for negative values, applying schema-specific anchor constraints for error-prone categories (e.g., vomiting, pain, Glasgow Coma Scale), requiring the presence of domain-specific lexical cues. ensuring values conform exactly to the allowed schema enumeration when applicable. This agent significantly reduces hallucinated and semantically invalid predictions while preserving correct extractions supported by explicit textual evidence.

3.3.4. Filtering Agent

Despite strict validation, development analysis revealed that certain observations remained over-predicted due to subtle semantic drift or overly permissive extraction. To address this, a Precision oriented Filter Agent is applied as an additional refinement stage. The precision filter operates at the level of individual observations. For each validated observation, the agent re-evaluates whether the extraction should be retained or discarded by considering the full transcript context, the schema definition and allowed values, and the extracted evidence span and value. The agent produces a binary decision (KEEP or DROP) for each observation, along with a brief rationale. The precision filter is intentionally conservative and is configured with deterministic decoding settings (temperature = 0.0) to ensure reproducibility. Its role is not to introduce new information, but to eliminate residual false positives that survive earlier validation stages.

Prompt Design Across Agents The prompting strategy differs across agents but follows a consistent evidence-grounded design. The extractor agent employs a schema-aware prompt that enforces strict constraints, requiring verbatim evidence spans and explicitly prohibiting inference or speculative reasoning. The precision filter operates using a binary decision prompt, where each candidate observation is evaluated independently and labeled as KEEP or DROP based on the schema definition, extracted value, evidence span, and transcript context. In contrast, the validation and suppression stages are fully rule-based and do not rely on prompting.

3.3.5. Conservative Suppression Table

In addition to agent-based filtering, a lightweight deterministic suppression mechanism is applied. The suppression table is implemented as a rule-based post-processing step derived from systematic error analysis on the development set. It is constructed by identifying observation categories that were consistently over-predicted, semantically redundant, or weakly grounded in the input text.

The suppression rules are divided into two groups: (i) an always-suppressed set, where specific observation categories are removed regardless of context due to persistent false-positive behavior (e.g., high-level or abstract clinical concepts), and (ii) a negation-sensitive set, where negative values (e.g., “No”, “Absent”) are retained only if explicit negation cues (such as “no”, “denies”, “without”, or “absent”) are present in the evidence span.

This mechanism is intentionally conservative and does not introduce any new predictions; it only removes unreliable ones. It is applied uniformly across development and test data without any test-time tuning, ensuring consistent behavior and reducing systematic false positives while minimizing the risk of discarding valid observations.

3.4. Model Selection

For all experiments, we selected LLaMA-3.3, a 70B-parameter decoder-only transformer model, as the core language model in our multi-agent pipeline. This choice was informed by results from our prior study, where the model demonstrated strong instruction-following abilities and robust performance on structured information extraction tasks. Importantly, we used the model strictly in inference-only mode, with no fine-tuning or task-specific adaptation on the MEDIQA-SYNUR dataset.

3.5. Implementation Details

All experiments were conducted using LLaMA-3.3 via the Ollama framework, enabling fully local inference to satisfy clinical data privacy constraints. In early development stages, we also evaluated DeepSeek as the underlying language model; however, due to unstable performance and significantly lower recall, all subsequent experiments were carried out using LLaMA-3.3. The extraction pipeline follows a multi-agent design, where transcript segmentation is applied to divide long clinical notes into manageable chunks. Each segment is processed independently by the extraction agent, which iteratively extracts candidate observations across schema batches.

The prompting strategy differs across agents but follows a consistent evidence-grounded design. The extractor agent uses a schema-aware prompt

that includes a subset of observation categories and instructs the model to output only observations explicitly supported by verbatim text spans from the transcript. The prompt enforces strict constraints, including prohibiting implicit inference, speculative reasoning, and requiring exact substring evidence. Negative values are only allowed when explicit negation cues are present in the transcript.

The validation stage is fully rule-based and does not rely on prompting. It applies deterministic checks to ensure that each extracted observation is supported by exact evidence in the transcript, does not contain hedging language, satisfies negation constraints, and conforms to schema-specific requirements, including value types and anchor-based rules for selected concepts.

The precision filtering agent operates as a final refinement step and uses a separate prompt to re-evaluate each observation individually. Given the observation, schema definition, and full transcript, the model produces a strict binary decision (KEEP or DROP). This stage is configured with deterministic decoding (temperature = 0.0) to ensure reproducibility and is designed to remove residual false positives that pass earlier validation stages.

Schema handling is performed through static schema segmentation, where the full set of observation categories is divided into fixed batches, and each batch is processed independently. This approach enables scalable processing of the full schema without exceeding prompt limitations while maintaining full coverage across all observation categories.

In addition to agent-based filtering, a deterministic suppression mechanism is applied as a lightweight post-processing step. The suppression table is derived from systematic error analysis on the development set and includes two categories: (i) always-suppressed observations that were consistently over-predicted or semantically redundant, and (ii) negation-sensitive observations, where negative values are retained only if explicit negation cues are present in the evidence.

To assess the contribution of each component in the extraction pipeline, we conducted a series of controlled experiments by selectively enabling or disabling key modules, including transcript segmentation (for handling long transcripts), the suppression table (to reduce systematic false positives), and the precision agent (which applies an additional LLM-based refinement step).

In addition, we explored alternative strategies, such as embedding-based schema matching, where schema items are aligned with transcript representations to guide extraction under varying relevance thresholds. However, these approaches were evaluated only as part of the experimental analysis and were not included in the final system configuration. This systematic evaluation allows us to quantify how each module, both individually and in combination, impacts overall extraction performance.

4. Error Analysis

To better understand the strengths and limitations of the proposed multi-agent clinical information extraction system, we conducted a comprehensive error analysis on the development set. Since ground-truth annotations are hidden during the test phase, all qualitative and quantitative analyses were performed using false positive (FP) and false negative (FN) predictions observed on the development data. The objective of this analysis is not only to quantify errors, but also to identify systematic failure modes and assess how different components of the pipeline contribute to precision–recall trade-offs.

4.1. False Positive Analysis

Figure 2 presents the top 30 observation categories most frequently associated with false positive predictions. Several high-frequency false positives correspond to semantically broad or overlapping clinical concepts, including gastrointestinal symptoms, gait and transferring, urinary symptoms, oxygen intake, and dyspnea. These categories often appear in narrative nursing documentation in descriptive or contextual forms, which increases the likelihood of over-extraction even under strict evidence constraints.

In many cases, false positives arose when: (i) the model extracted implicit or contextual mentions rather than explicitly stated observations; (ii) multiple related concepts were mentioned in proximity (e.g., mobility, transfer technique, gait), leading to semantic boundary confusion; and (iii) measurement-related concepts (e.g., pulse oximetry unit, MAP unit) were mentioned as part of documentation templates rather than clinical findings.

These findings motivated the introduction of rule-based suppression tables and a precision filter agent, specifically targeting categories with historically high FP rates.

4.2. False Negative Analysis

Figure 3 shows the top 30 observation categories most frequently missed by the system (false negatives). False negatives were most prominent in categories such as orientation, secondary diagnosis, cognitive status, skin condition, and dyspnea. These concepts often exhibit: (i) high lexical variability (e.g., multiple ways of describing cognitive or mental status); (ii) distributed mentions across the transcript rather than a single explicit phrase; and (iii) subtle negation or partial descriptions that are

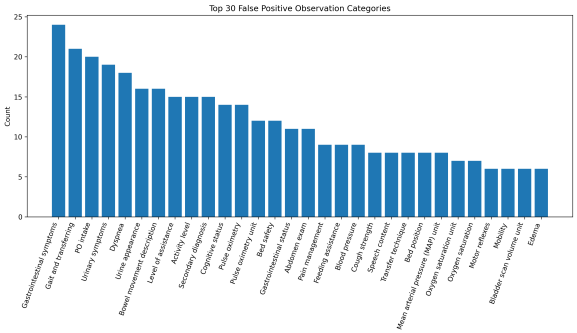


Figure 2: Top 30 observation categories most frequently associated with false positive predictions in the development set.

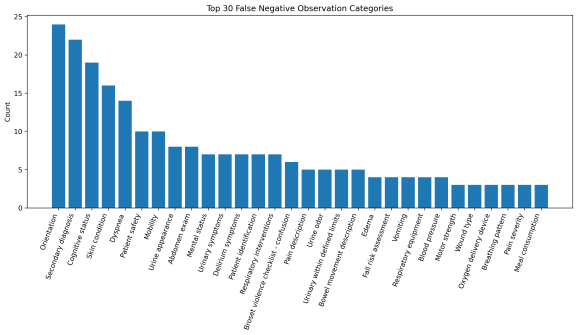


Figure 3: Top 30 observation categories most frequently associated with false negative predictions in the development set.

difficult to capture under strict no-inference rules. Additionally, some false negatives resulted from conservative validation and suppression strategies, where ambiguous or weakly anchored extractions were intentionally discarded to preserve precision.

4.3. Error Cause Distribution

To further characterize system behavior, we grouped all FP and FN instances into high-level error cause categories, summarized in Figure 4. The dominant error patterns include:

(i) **Other FP causes**: the largest share of false positives. This category captures residual errors not covered by predefined rule-based categories, and primarily arises from three sources: (a) *semantic ambiguity*, where phrases partially match a schema concept but lack sufficient specificity (e.g., vague mentions of discomfort interpreted as pain); (b) *template-driven artifacts*, where structured clinical templates or repeated phrasing trigger spurious matches without true clinical evidence; and (c) *borderline evidence spans*, where extracted text technically satisfies verbatim matching constraints but does not strongly support the assigned observation value. These errors reflect the limitations of strict substring-based evidence matching when applied

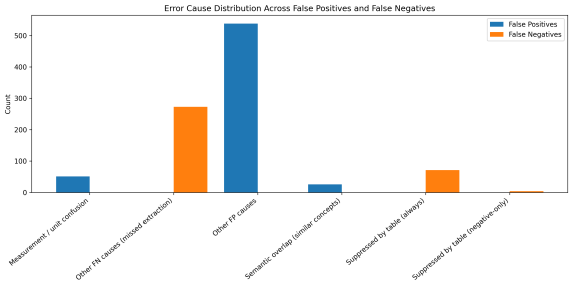


Figure 4: Distribution of high-level error causes for false positive and false negative predictions, aggregated over all development records.

to loosely structured clinical language.

(ii) **Other FN causes (missed extraction)**: the most common false negative category, typically due to strict evidence requirements or insufficient lexical anchoring;

(iii) **Measurement / unit confusion**: errors where numeric values or units were present but not clearly attributable to a specific observation category;

(iv) **Semantic overlap (similar concepts)**: errors arising from closely related schema categories (e.g., mental status vs. orientation vs. delirium symptoms); and

(v) **Suppression-related effects**: a smaller subset of false negatives attributable to conservative suppression rules, particularly for negative-only observations.

Importantly, these trends reflect an intentional precision-oriented design trade-off, where the system favors high-confidence, explicitly grounded extractions over recall in ambiguous cases.

5. Results

All results reported in this section were obtained before the official competition deadline No additional post-deadline experiments were conducted.

5.1. Development Set Exploration

All development-set experiments were conducted through extensive exploratory trial-and-error to better understand the behavior of the extraction pipeline. Our initial trials employed DeepSeek as the underlying model. Although the precision-filtered variant produced a high precision of 0.6984, its recall was extremely low (0.0633), resulting in an F1 score of only 0.1161. Efforts to improve recall by injecting internal evidence into the extraction process further degraded performance, with recall dropping to 0.0115 and the F1 score to 0.0224. These findings indicated that DeepSeek could not generalize well to this task, particularly in retrieving

relevant information from long and detailed clinical transcripts. Consequently, DeepSeek was excluded early in development, and all subsequent experiments were carried out using LLaMA 3.3.

Before settling on a stable configuration, we also examined a rule-based evidence–augmented pipeline, which achieved a very high recall of 0.7978 but substantially reduced precision, yielding an overall F1 score of 0.6039. With LLaMA 3.3 as the foundation, system refinement became significantly more effective. Incorporating a disambiguation agent alongside the precision filter and suppression table produced a notable improvement, reaching a precision of 0.6268, a recall of 0.7201, and an F1 score of 0.6702. This phase highlighted the importance of resolving ambiguous or overlapping observations in the transcripts. Further ablation studies demonstrated that both the suppression table derived from systematic error analysis, and the precision-filtering step were essential to maintaining balanced performance. When either component was removed, the system's F1 score dropped sharply to 0.1902, mainly due to a substantial decline in recall.

The decisive improvement occurred when transcript segmentation was integrated with the precision agent and the suppression table within the LLaMA 3.3 pipeline. This configuration yielded the strongest development-set performance, with a precision of 0.6427, a recall of 0.7518, and an F1 score of 0.6930. As illustrated in Figure 5, this setting demonstrated the most stable and balanced performance among all pre-deadline experiments and was therefore adopted as the final configuration for the system.

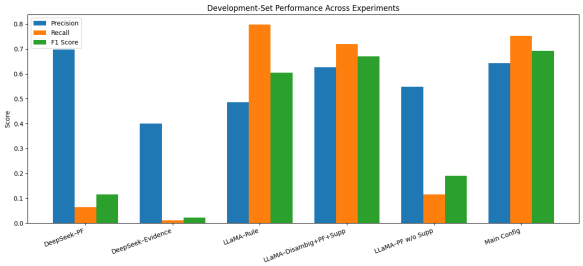


Figure 5: Development-set performance across major experimental configurations. "DeepSeek–PF" and "DeepSeek–Evidence" represent early baselines using the DeepSeek model with precision filtering or evidence-enriched prompting. "LLaMA–Rule" denotes rule-based evidence extraction with LLaMA 3.3. "LLaMA–Disambig+PF+Supp" includes the disambiguation agent, precision filtering, and the suppression table. "LLaMA–PF w/o Supp" shows the effect of disabling the suppression table. "Main Config" represents the final development-set configuration selected for test-set evaluation.

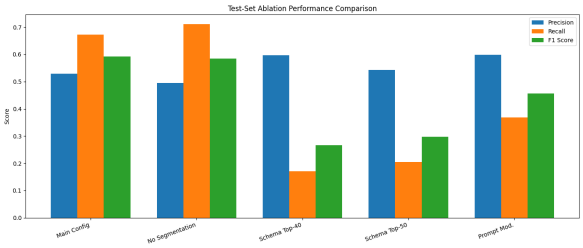


Figure 6: Test-set performance of the main configuration compared with ablation variants. "Main Config" refers to the full system using transcript segmentation, the precision agent, and the suppression table. "No Segmentation" removes transcript segmentation while retaining the precision agent. "Schema Top-40" and "Schema Top-50" restrict extraction to the top schema-matched concepts using embedding similarity. "Prompt Mod." applies prompting adjustments without segmentation or the suppression table. The main configuration achieves the best overall F1 score on the test set.

5.2. Test Set Results

The system that achieved the highest score on the development set was carried forward and evaluated directly on the test set. This configuration, combining transcript segmentation, the precision agent, and the suppression table, served as our primary setup. On the test set, it achieved precision = 0.5292, recall = 0.6725, and F1 = 0.5923 (ref_obs = 2721; hyp_obs = 3458), confirming its strong and balanced performance across metrics.

After establishing this main system, we conducted a series of additional experiments in an effort to further improve extraction quality. These included removing segmentation, enabling or disabling the precision agent, experimenting with the suppression table, and introducing schema-guided extraction using embedding similarity with different schema subset sizes (top-40, top-50, and top-70 items).

However, none of these enhancement attempts outperformed the main configuration. As shown in Figure 6, disabling segmentation while keeping the precision agent active resulted in a slightly lower F1 score of 0.5838. Schema-based variants, despite achieving high precision in some cases, suffered substantial recall losses, leading to significantly lower F1 scores (e.g., 0.2662 with top-40 items and 0.2969 with top-50 items). Even the broader schema variant with extraction prompt modification, tested without segmentation, with precision agent, without suppression table, reached only F1 = 0.4560.

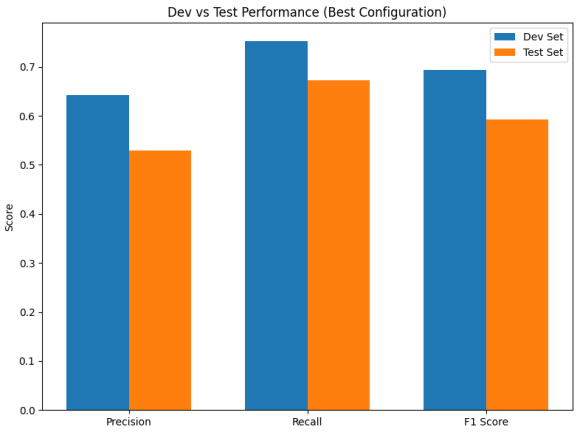


Figure 7: Comparison of development-set and test-set performance using the best configuration of the proposed system. The model integrates transcript segmentation, the precision agent, and the suppression table, which together achieved the highest score during development. Performance is shown across Precision, Recall, and F1 Score. While the test set shows a consistent drop across all metrics, as expected due to dataset shift, the configuration still maintains strong generalization, confirming it as the optimal setting for final evaluation.

5.3. Main Results: Best Configuration

After extensive exploration, the final configuration submitted to the competition consisted of Transcript segmentation (enabled), Precision agent (enabled), and Suppression table (enabled and derived from dev-set analysis). This setup achieved strong performance on the development set, with an F1 score of 0.6930 (precision = 0.6427, recall = 0.7518). The same configuration was then applied directly to the test set, without any further changes, resulting in an F1 score of 0.5923 (precision = 0.5292, recall = 0.6725).

As shown in Figure 7, these results reflect the final system state as it existed before the competition deadline, and they represent the most effective balance of precision and recall identified through our iterative exploration.

6. Discussion

The experiments conducted prior to the competition deadline highlight the core strengths and limitations of the proposed extraction pipeline. Overall, the system demonstrated strong recall, driven largely by effective transcript segmentation and the combined influence of the precision agent and suppression table. Segmentation proved especially valuable for isolating clinically dense sections of text, allowing the model to capture relevant observations consistently and with fewer omissions. However, the pipeline also exhibited persistent precision challenges due to the generative nature of large language models. Despite explicit instructions to avoid inference, the model sometimes produced clinically plausible but unsupported observations, a form of generative drift. This tendency to "over-interpret" clinical cues resulted in hallucinated predictions such as inferred nutrition status or IV fluid details that were not present in the transcript. The example analysis further illustrated how the model could accurately capture ground-truth observations while simultaneously introducing additional, schema-shaped predictions without clear evidence. To counteract these issues, the precision agent and suppression table played a crucial role. The precision agent filtered weakly supported predictions, while the suppression table blocked patterns of recurrent hallucinations identified during error analysis. Ablation results showed that removing either component substantially increased false positives, confirming their necessity for maintaining competitive precision and overall F1 performance. In summary, the pipeline's success stemmed from its ability to balance high recall with post-processing mechanisms that restrained the generative biases of LLMs. Nonetheless, the observed drift underscores ongoing challenges in evidence-grounded extraction and highlights areas for further refinement.

7. Conclusion

In this work, we presented a multi-agent LLM-based system for structured nursing observation extraction from clinical transcripts, addressing the challenges of evidence grounding and value normalization. By decomposing the extraction process into specialized agents and iteratively refining the pipeline through systematic error analysis, our system achieved a strong balance of precision and recall, with an F1 score of 0.6930 on the development set and 0.5923 on the test set. Despite these results, limitations remain, particularly the reliance on hand-crafted suppression tables and transcript-specific processing, which can reduce robustness across domains. Future work will explore fine-tuning the underlying LLM on synthetic or multi-institutional clinical transcripts, improving model-level controllability, and incorporating human-in-the-loop validation to create a more reliable and generalizable system. Overall, our findings demonstrate that a multi-agent LLM approach can effectively structure free-text clinical documentation, providing a foundation for more automated and evidence-grounded clinical data extraction.

8. Limitations

Despite the strong performance of the proposed system, several limitations must be acknowledged. First, although prompting strategies and suppression mechanisms were designed to strictly constrain the model to evidence-based extraction, the underlying LLM still exhibited generative tendencies. This occasionally led to clinically plausible but unsupported predictions, lowering precision, particularly for schema fields that were semantically broad or frequently referenced in the transcripts. Notably, the suppression table, crafted based on the development set, was also used during testing; this resulted in reduced performance, highlighting the sensitivity of static, hand-crafted controls. Second, the system relies heavily on transcript segmentation, precision filtering, and the manually derived suppression table. While effective, these components require substantial engineering effort and are sensitive to variations in transcript style or domain-specific phrasing.

9. Bibliographical References

Asma Ben Abacha, Yassine M'rabet, 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.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th bioNLP workshop and shared task*, pages 370–379.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen-Yildiz. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513.

Hammaad Adam, Junjing Lin, Jianchang Lin, Hillary Keenan, Ashia Wilson, and Marzyeh Ghassemi. 2025. Clinical information extraction with large language models: A case study on organ procurement.

Ragnhildur I Bjarnadottir, Walter O Bockting, Sunmoo Yoon, and Dawn W Dowding. Computers, informatics, nursing nurse documentation of sexual orientation and gender identity in home healthcare: A text mining study-manuscript draft-manuscript number: Cin-d-17-00098r1 full title: Nurse documentation of sexual orientation and gender identity in home healthcare: A text mining study powered by editorial manager® and produxion manager® from aries systems corporation. Technical report.

Ragnhildur I. Bjarnadottir and Robert J. Lucero. 2018. What can we learn about fall risk factors from ehr nursing notes? a text mining study. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 6:21.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jérôme Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, et al. 2025a. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 859–870.

Jean-Philippe Corbeil, Asma Ben Abacha, Jérôme Tremblay, Phillip Swazinna, Akila Jeeson Daniel, Miguel Del-Agua, and François Beaulieu. 2025b. Overview of the mediqa-oe 2025 shared task on medical order extraction from doctor-patient consultations. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 11–16.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025c. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 859–870, Suzhou (China). Association for Computational Linguistics.

Cari Bader Nate Bodenstab Asma Ben Abacha George Michalopoulos, Jean-Philippe Corbeil. 2026. Overview of the mediqa-synur 2026 shared task on observation extraction from nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

Adi V. Gundlapalli, Guy Divita, Andrew Redd, Marjorie E. Carter, Danette Ko, Michael Rubin, Matthew Samore, Judith Strymish, Sarah Krein,

Kalpana Gupta, Anne Sales, and Barbara W. Trautner. 2017. Detecting the presence of an indwelling urinary catheter and urinary symptoms in hospitalized patients using natural language processing. *Journal of Biomedical Informatics*, 71:S39–S45.

Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R. Pisani, and Kathryn Turner. 2023. Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review.

Kexin Huang, Tamryn F. Gray, Santiago Romero-Brufau, James A. Tulsky, and Charlotta Lindvall. 2021. Using nursing notes to improve clinical outcome prediction in intensive care patients: A retrospective cohort study. *Journal of the American Medical Informatics Association*, 28:1660–1666.

Sookyung Hyun, Stephen B. Johnson, and Suzanne Bakken. 2009. Exploring the ability of natural language processing to extract data from nursing narratives. *CIN - Computers Informatics Nursing*, 27:215–223.

Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F. Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review.

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, 17(01):128–144.

Shazia Mitha, Jessica Schwartz, Mollie Hobensack, Kenrick Cato, Kyungmi Woo, Arlene Smaldone, and Maxim Topaz. 2023. Natural language processing of nursing notes: An integrative review.

Nadia Saeed. 2024. Medifact at mediqa-m3g 2024: Medical question answering in dermatology with multimodal learning. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 339–345.

Danielle Scharp, Mollie Hobensack, Anahita Davoudi, and Maxim Topaz. 2024. Natural language processing applied to clinical documentation in post-acute care settings: A scoping review.

Lei Wang, Yinyao Ma, Wenshuai Bi, Hanlin Lv, and Yuxiang Li. 2024. An entity extraction pipeline for medical text records using large language models: Analytical study. *Journal of Medical Internet Research*, 26.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review.

Isabella Catharina Wiest, Dyke Ferber, Jiefu Zhu, Marko van Treeck, Sonja K. Meyer, Radhika Juglan, Zunamys I. Carrero, Daniel Paech, Jens Kleesiek, Matthias P. Ebert, Daniel Truhn, and Jakob Nikolas Kather. 2024. Privacy-preserving large language models for structured medical information retrieval. *npj Digital Medicine*, 7.

A. Additional Dataset Analysis

This study uses the dataset released as part of the MEDIQA-SYNUR shared task, hosted on the Codabench evaluation platform. The task focuses on extracting structured nursing observations from unstructured clinical transcripts under strict schema and evidence constraints. Each record consists of a free-text nursing transcript paired with gold-standard structured annotations. The dataset is divided into training, development, and test splits. The training split contains 122 patient records, while the development split contains 101 patient records. The test split contains 199 patient records. Analysis of concept distributions across the training and development splits shows consistent usage patterns, suggesting that the data is reasonably balanced at the split level despite intrinsic variability across clinical observations. The transcripts follow a conversational clinical style and frequently contain hesitations, discourse markers, and loosely structured descriptions, characteristics also noted in the original dataset publication. This introduces realistic challenges for automated extraction systems, including ambiguous phrasing, overlapping concepts, and implicit contextual cues.

Figures 8 and 9 illustrate the top 30 most frequent nursing observation categories in the training and development sets, respectively. Across both splits, commonly annotated observations include cognitive status, mobility, oxygen delivery device, oxygen saturation, urinary symptoms, and skin turgor. In contrast, many observation categories occur infrequently, reflecting the broad range of clinical scenarios and variability in nursing documentation captured by the dataset. The dataset exhibits a pronounced long-tailed distribution, in which a small subset of observation categories accounts for a

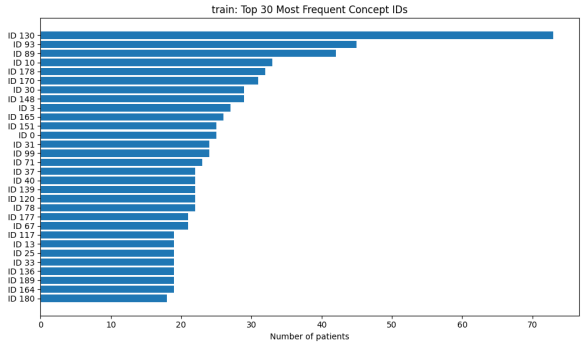


Figure 8: Top 30 most frequent observation concept IDs in the training set. The distribution highlights a long-tailed pattern, where a small number of observation categories occur frequently across patients, while many schema concepts appear sparsely.

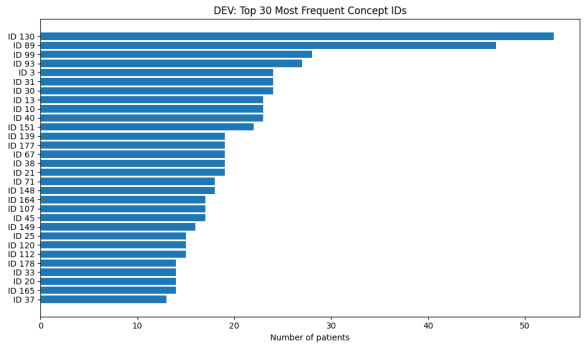


Figure 9: Top 30 most frequent observation concept IDs in the development set. The distribution highlights a long-tailed pattern, where a small number of observation categories occur frequently across patients, while many schema concepts appear sparsely.

large proportion of annotations, while a substantial number of categories appear only sparsely. This distribution poses two key challenges for automatic extraction systems. First, achieving high recall for rare observation categories is difficult due to limited training signals. Second, maintaining high precision for frequently occurring but semantically broad categories is challenging, as such categories often encompass overlapping or closely related clinical concepts. In addition, semantic overlap among related observations, such as cognitive status, orientation, and delirium symptoms, further increases task complexity and contributes to potential ambiguity during extraction. Importantly, all gold-standard annotations are strictly grounded in the transcript text. The task requires systems to extract only observations that are explicitly stated and to provide verbatim evidence spans from the transcript, without relying on inference or clinical interpretation beyond the documented text.

LTRC-IIIT at MEDIQA-SYNUR 2026: Benchmarking a Fully Local, Training-Free RAG Pipeline

Aashwin Vaish, Dipti Misra Sharma
Language Technologies Research Center, KCIS
IIIT Hyderabad, India
aashwin.vaish@research.iiit.ac.in, dipti@iiit.ac.in

Abstract

In this paper we present our solution to MEDIQA-SYNUR 2026 shared task organized at LREC-ClinicalNLP workshop. The goal of the task is to populate Electronic Health Record (EHR) flowsheets using the transcriptions of nurse dictations, to alleviate the extensive manual labor associated with sifting through large flowsheets of clinical concepts. We propose a modular architecture combining heuristic-driven Retrieval-Augmented Generation (RAG) with grammar-constrained decoding on an open-weight, quantized, 8B-parameter model (Llama 3.1 Instruct). Our system achieves an F1 score of 0.57, significantly trailing the initial zero-shot experiments with GPT-4o (F1 $\approx$ 0.88) and placing it towards the lower end of the current leaderboard. We conduct a failure analysis of this approach while establishing a baseline for privacy-preserving, zero-shot documentation assistants.

Keywords: Clinical Documentation, Local LLMs, Zero-Shot Extraction, Data Privacy, Negative Results

1. Introduction

While the digitization of healthcare has expedited several aspects of clinical care, it has also placed additional burdens of documentation on the clinicians and the nurses (Budd, 2023). One such documentation task involves filling Electronic Health Records (EHR) flowsheets that record aspects of patient care like vitals, doctor's assessments, intake/output levels etc. These flowsheets often span multiple tabs with hundreds of ontological concepts that need to be sifted through to find the relevant fields. Naturally, automation of this task by populating the EHR flowsheet using nurse dictations will improve the time that nurses can devote to patient care. To encourage research on this front, the MEDIQA-SYNUR shared task was proposed (George Michalopoulos, 2026).

While the recent advances of commercially deployed LLMs including the likes of OpenAI's GPT and Google's Gemini have shown impressive results on information extraction in healthcare (Umeton et al., 2023), such solutions have increased concerns of data privacy and sovereignty over local open-weight solutions.

Furthermore, due to limited availability of public patient data for training on this task, any comprehensive LLM training task incurs a significant cost. So the question we ask is -

How effectively can we automate flowsheet extraction using only local, general-purpose resources without task-specific training?

We present a training free pipeline designed to run on consumer hardware. Our approach utilizes a standard retrieval architecture (Lewis et al., 2020) enhanced by heuristics, decomposing enumerable fields into individual vectors for retrieval and using deterministic rules to link dependent variables (e.g., units of measurement). For generation, we employ Llama 3.1 (8B Instruct)(Dubey et al., 2024) with grammar-constrained decoding to enforce schema compliance without fine-tuning. Our results yield an F1 score of 0.57, placing our method at the lower end of the current leaderboard. However, this lower performance comes with distinct operational advantages:

1. **Privacy:** Data never leaves the local machine.
2. **Adaptability:** The system is purely zero-shot, handling a new flowsheet requires only updating the JSON schema, not retraining.
3. **Accessibility:** The pipeline runs on standard hardware, lowering the barrier to entry for resource-constrained institutions.

In this work, we detail our heuristic retrieval strategies and provide a transparent error analysis, identifying the specific reasoning gaps in smaller scale LLMs that currently prevent them from matching supervised performance in the medical domain.

2. Methodology

Our pipeline consists of two distinct stages: a retrieval module designed to filter relevant flowsheet rows from a potential set of hundreds, and a generation module that extracts values using a local instruction LLM. The architecture prioritizes recall in the retrieval stage and precision in the generation stage, operating entirely on consumer-grade hardware (i.e., a single NVIDIA GPU).

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 136–140
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

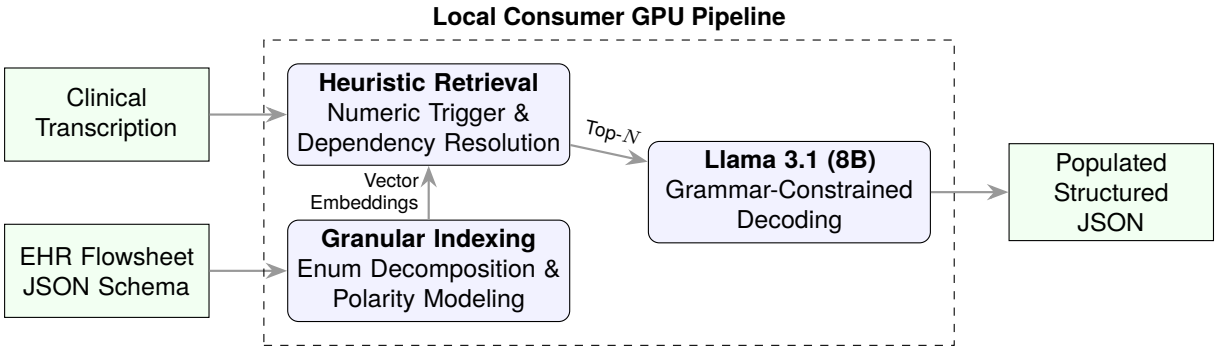


Figure 1: Architecture of the zero-shot, training-free RAG pipeline. The system operates entirely locally, utilizing heuristic-augmented retrieval to filter the JSON schema before passing candidates to the quantized 8B-parameter LLM.

2.1. Granular State Indexing

The easiest approach to make an RAG system would be to embed entire database rows as single vectors. In the context of flowsheet schemas, this obscures a critical semantic distinctions. For example, a row defining "Weightbearing Status" might contain the Enum options *["Full", "Partial", "Non-Weightbearing"]*. Embedding the row metadata alone fails to capture the semantic distance between these mutually exclusive states. To address this, we implement a state indexing strategy. Instead of a single vector per row, we generate multiple vectors based on the field type:

- **Categorical Fields:** We decompose Enum options into individual "child" vectors (e.g., *"Weightbearing: Non-Weightbearing"*). This allows the retriever to match specific patient states rather than generic field names.
- **Boolean Polarity:** For binary fields (Yes/No), we explicitly model polarity. Rather than embedding generic "Yes" and "No" tokens, which introduce noise, we generate positive assertions (e.g., *"Edema present"*) and negative assertions (e.g., *"No Edema"*).

All child vectors map back to their original parent Row ID, allowing us to retrieve the full schema definition if any specific state is detected.

2.2. Heuristic Enhanced Retrieval

Clinical transcriptions often have high information density and a disjointed syntax, along with disfluencies. After splitting on common disfluencies (e.g., "um", "uh"), we apply a sliding window chunking strategy, splitting text on commas rather than sentences. This reduces vector dilution, ensuring that short, distinct observations (e.g., *"lungs clear"*, *"heart regular"*) are represented by sharp, high-magnitude vectors.

Using the *sentence-transformers/all-MiniLM-L6-v2* embedding model, we augment the standard cosine similarity search (retrieving the top $N = 15$ candidates) with two deterministic heuristics to address common vector space failures:

2.2.1. Type Aware Retrieval

Vector models frequently fail to associate raw numbers (e.g., *"88"*, *"38.5"*) with their corresponding clinical concepts (e.g., *Heart Rate*, *Temperature*). We implement a **Numeric Trigger**: if a text chunk contains a digit, the system executes a parallel search restricted solely to fields defined as `NUMERIC` in the schema. The top-scoring numeric field is forced into the candidate list, prioritizing type compatibility over semantic similarity.

2.2.2. Deterministic Dependency Resolution

Certain fields, such as units of measurement (e.g., *Temperature Unit*), are rarely explicitly vocalized by clinicians but are structurally required by the EHR. Pure vector search often fails to retrieve these dependent variables. We address this via **Dependency Resolution**: if a primary numeric field is retrieved (e.g., *Temperature*), its associated metadata fields defined in the schema are automatically injected into the context window, bypassing the retrieval process entirely.

2.3. Constrained Generative Extraction

The final extraction is performed by **Llama-3.1-8B-Instruct**, quantized to 8-bit precision. While 8B-parameter models are generally capable of zero-shot extraction, they are prone to outputting invalid JSON and inventing values for unmentioned fields.

We mitigate these issues through **Grammar-Constrained Decoding**. We utilize the Grammar-Based Network Format (GBNF) engine within the `llama.cpp` serving infrastructure. Rather than relying on the LLM to learn JSON syntax through

prompt examples, the GBNF engine dynamically masks the model's output logits during inference. It assigns a zero probability to any token that violates the provided JSON schema. This strictly limits generation to valid brackets, schema-defined keys (e.g., `id` and `value`), and appropriate string enclosures, ensuring 100% syntactical validity without fine-tuning.

To control hallucination, we employ a strict "Silence Constraint" in the system prompt, instructing the model to output a null value if a field is not explicitly supported by the text. This shifts the burden of decision-making from implicit inference to explicit extraction.

2.3.1. Generation Prompt

```
You are a strict data extraction
engine.
RULES:
1. Output specific JSON only.
2. SILENCE CONSTRAINT: If a field
from the schema is NOT mentioned,
do NOT include it.
3. Select value(s) ONLY if they
appear in text or are HEAVILY
implied by BASIC LOGIC.
4. FORMAT: List of objects with
"id" and "value". All values must
be strings.

CONTEXT:
{transcription}

TARGET SCHEMA:
{candidates}
```

2.4. Deterministic Post-Processing

Despite strict JSON structural constraints, the generative model occasionally exhibits semantic slippage by paraphrasing rigid `SINGLE_SELECT` and `MULTI_SELECT` Enum values (e.g., generating *"weightbearing"* instead of the required *"weightbearing as tolerated"*). Because downstream EHR databases require exact string matches, we implement a two-step deterministic post-processing layer to sanitize the LLM outputs.

The system uses a localized vector-similarity check. We embed both the LLM's predicted string and all valid Enum options using the same *all-MiniLM-L6-v2* model utilized in the retrieval stage. The predicted value is then deterministically mapped to the valid option exhibiting the highest cosine similarity.

To prevent aggressive misclassification of genuine hallucinations, we enforce a strict minimum similarity threshold. If the highest-scoring option falls below this threshold (<0.6), the extracted value is discarded. This operation runs in $O(1)$ time relative to the database size, as it only embeds the isolated schema options rather than querying the global vector store, thereby correcting near-miss extractions without adding significant computation overhead.

3. Experiments and Results

3.1. Evaluation Setup

We evaluated our local-first extraction pipeline against the MEDIQA-SYNUR dataset (Corbeil et al., 2025). The dataset contains a concept schema in a JSON format that specifies a mixture of `NUMERIC`, `STRING`, `SINGLE_SELECT`, and `MULTI_SELECT` variables. We measured the system performance using the shared task evaluation script. Additionally, we measured performance of the retrieval step independently from the generation step to better analyze the impact of either step.

3.2. Retrieval-Stage Performance

To isolate the effectiveness of our heuristic-augmented vector search, we evaluated the Retrieval-Augmented Generation (RAG) filter independently of the downstream LLM. The heuristic based retrieval module achieved a Precision of 0.35, a Recall of 0.87, and an F1 score of 0.50. This high-recall, low-precision profile is an intentional architectural trade-off. In the context of EHR flowsheet extraction, missing a relevant clinical observation (a false negative) at the retrieval stage is a catastrophic failure, as the downstream LLM should not extract data it is never shown. But on the other hand, an irrelevant schema row (a false positive) merely consumes additional context window tokens.

By setting generous retrieval limits and applying our "Numeric Boost" heuristic, the retrieval system successfully surfaces 87% of the necessary schema context required to process the transcription. The burden of precision is then intentionally offloaded to the generative stage. The LLM acts as a secondary precision filter, using its semantic understanding and strict "Silence Constraints" to ignore the irrelevant candidate rows (the 65% noise generated by the 0.35 precision rate), ultimately raising the end-to-end system precision to 0.64.

3.3. Baseline Performance

Our zero-shot, locally hosted pipeline achieved a Precision of 0.6, a Recall of 0.55, and an F1 score of 0.57. For context, the current best score on the shared task leaderboard (F1 = 0.84) was around the same as the prelimnary experiments conducted by the organizers using GPT-4o, which achieved

an F1 score of approximately 0.88 in zero shot setting (Corbeil et al., 2025). While our system trails the organizers baseline by a significant margin (Table 1), it establishes a functional baseline for a fully zero-training, privacy-preserving architecture. More importantly, analyzing the delta between our 0.57 F1 and the 0.84 baseline reveals specific limitations of using compact LLMs for deterministic extraction.

3.4. Failure Analysis

To understand the performance gap, we conducted a manual qualitative analysis of the 204 missed occurrences in our validation set (the test set references had not been released at the time of writing). The errors largely fell into three distinct categories inherent to compact LLMs:

1. The semantic gap in abstract strings: Of the missed variables, a significant portion were generic `STRING` fields, specifically abstract categories like *"Secondary Diagnosis"*.

While the RAG filter easily retrieved concrete fields (e.g., *"Heart Rate"*), it struggled to map specific diseases mentioned in the text (e.g., *"Sepsis"*) to the abstract schema field without explicit anchor examples in the vector space.

2. Format slippage and enum paraphrasing: Despite grammar-constrained decoding ensuring valid JSON, the 8B-parameter model frequently paraphrased strict Enum values.

For example, generating *"weightbearing"* instead of the required *"weightbearing as tolerated"*.

While our deterministic post-processing layer (Section 2.4) successfully mitigated the majority of these paraphrasing errors by mapping them back to valid Enums via cosine similarity, extreme deviations (e.g. an entire sentence being picked from the dictation) or highly truncated generations occasionally fell below our safety threshold and remained penalized as false negatives.

3. The "Silence Constraint" Failure: LLMs exhibit a strong completion bias (Kumar et al., 2025). When presented with a flowsheet containing binary fields (e.g., *"Violence Checklist: [Yes, No]"*), the model feels compelled to select an option even if the clinical transcript is completely silent on the matter. Despite explicit prompt instructions to omit unmentioned fields, the model occasionally hallucinated *"No"* or *"None"*. This contributed heavily to precision degradation.

3.5. Throughput and Hardware Efficiency

While trailing in accuracy, our system excelled in operational efficiency. By implementing continuous batching with a shared context window across multiple execution slots, the entire pipeline (embedding generation, heuristic vector search, and 8-bit quantized LLM generation) ran satisfactorily on consumer-grade hardware. (NVIDIA 2080 Ti GPUs).

4. Conclusion and Future Work

In this work, we explored the feasibility of a fully local, zero-shot extraction pipeline for automating EHR flowsheet documentation. By combining granular state indexing and heuristic-augmented retrieval with grammar-constrained decoding on an 8B-parameter LLM, we established a privacy-preserving baseline that requires no task-specific fine-tuning or cloud API reliance. While our F1 score of 0.57 trails state-of-the-art closed-weight commercial AI systems, our failure analysis highlights the exact boundaries of compact LLMs in medical contexts: they excel at strict, explicit extraction but falter at implicit clinical reasoning and semantic gap bridging. However, the system's ability to run locally on consumer-grade hardware with high throughput (via continuous batching) makes it an operationally viable foundation for privacy-restricted environments. Future work will focus on narrowing the reasoning gap without sacrificing the zero-shot adaptability of the pipeline. We plan to explore multi-agent reflection loops, where a secondary local model verifies the logical consistency of the extracted flowsheet and the integration of lightweight, domain-specific knowledge graphs into the retrieval step to better anchor abstract clinical concepts (e.g., mapping specific diseases to "Secondary Diagnosis" fields).

Limitations

While this work establishes a baseline for local flowsheet extraction, several limitations must be acknowledged:

- **Performance Gap:** The 0.57 F1 score indicates that the system is not yet suitable for fully autonomous deployment. It is strictly an assistive technology requiring human verification.
- **Reasoning Constraints:** Our pipeline relies heavily on explicit mentions in the text. It struggles to infer values that a human clinician would consider "obvious" context (e.g., inferring that a patient walking unassisted implies a negative finding for fall risk).
- **Hardware Requirements:** Although technically "consumer-grade", our continuous batching setup utilizes two NVIDIA 2080 Ti GPUs (approx. 22GB VRAM). Leaving the possibility

System	Precision	Recall	F_1 score
Our Local Pipeline (Using Llama 3.1 8B)	0.60	0.55	0.57
GPT-4o Zero-Shot (Corbeil et al., 2025)	–	–	**0.88**

Table 1: Performance metrics of the retrieval stage, the full end-to-end local pipeline, and the commercial model GPT-4o zero-shot baseline.

for more aggressive quantization and serialization (as opposed to parallization) under further constraints. While vastly more accessible than the enterprise clusters required for 70B+ parameter models, this hardware footprint still exceeds the capacity of, say, individual practices, necessitating a dedicated local server architecture.

- **Language and Scope:** The pipeline was designed and evaluated exclusively on English-language clinical transcriptions and standard US-based EHR flowsheet schemas. Its robustness against heavily accented speech transcriptions or non-standard abbreviations remains untested due to the partially synthetic nature of the dataset.

Ethics Statement

The automation of medical documentation introduces significant ethical considerations, primarily concerning patient safety and data privacy.

Data Privacy: The primary motivation for this architecture is the protection of Protected Health Information (PHI). By utilizing a 100% local execution pipeline, patient data never traverses external networks or third-party APIs (e.g., OpenAI, Anthropic), ensuring strict compliance with data sovereignty regulations such as HIPAA and GDPR.

Patient Safety and Automation Bias: Errors in EHR flowsheets can lead to incorrect clinical decisions and patient harm. Because our system exhibits a known performance gap (F1 = 0.57) and compact LLMs are prone to occasional "hallucinations" (inventing positive or negative findings when none are mentioned), this system must strictly be deployed under a *Human-in-the-Loop* paradigm. The system is designed to draft flowsheet entries, not finalize them. Clinicians must review, edit, and authorize all generated structured data before it is committed to the patient's permanent medical record. Mitigating automation bias, where clinicians might blindly trust the AI's output to save time, is critical and requires careful UI/UX design in the final clinical application.

5. Bibliographical References

Jeffrey Budd. 2023. Burnout related to electronic health record use in primary care. *Journal of primary care & community health*, 14:21501319231166921.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Meta AI*.

Cari Bader Nate Bodenstab Asma Ben Abacha George Michalopoulos, Jean-Philippe Corbeil. 2026. Overview of the mediqa-synur 2026 shared task on observation extraction from nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

Charaka Vinayak Kumar, Ashok Urlana, Gopichand Kanumolu, Bala Mallikarjunarao Garlapati, and Pruthwik Mishra. 2025. No llm is free from bias: A comprehensive study of bias evaluation in large language models.

Patrick Lewis, Ethan Perez, Aleksandara 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.

Renato Umeton, Anne Kwok, Rahul Maurya, Domenic Leco, Naomi Lenane, Jennifer Willcox, Dana-Farber Committee, and Jason Johnson. 2023. Gpt-4 in a cancer center: Institute-wide deployment challenges and lessons learned.

BDI at MEDIQA-EVAL 2026: A ReAct-Style Multimodal Agent for Fine-Grained Medical Response Assessment

Justin Xu[1], Zizheng Zhang[1], Augustine Luk[1], Benjamin Khong[2]
Haochen Cui[1], Samuel Hwang[3], Alyssa Pradhan[1], Kevin Yuan[1], David W Eyre[1]
[1]University of Oxford, [2]University of California, Los Angeles, [3]University of California, Berkeley

Abstract

Free-text evaluation of multimodal clinical question answering (QA) systems remains a central challenge in medical NLP due to the complexity of medical knowledge, the necessity of integrating visual and textual information, and the limitations of existing automatic evaluation metrics for open-ended outputs. In this work, we present a training-free, agentic evaluation framework that formulates response scoring as evidence-guided orchestration of components rather than a task requiring conventional end-to-end fine-tuning of underlying LLMs/VLMs. Our ReAct-style evaluator combines (i) structured reasoning, (ii) multimodal retrieval of similar encounters, (iii) auxiliary explainable feature-based regression models that provide numeric priors and human-interpretable signals, (iv) VLM-generated visual QA references for comparison, and (v) optional image augmentation tools. Unlike standard LLM-as-a-judge approaches that rely on direct generative scoring, our agent decomposes evaluation into modular stages of evidence acquisition, structured feature modeling, and integrative reasoning. We apply this architecture to the MEDIQA-EVAL shared task – a multimodal, multilingual clinical evaluation challenge that assesses system-generated answers for patient queries paired with images along multiple clinical quality dimensions. We report results across both English and Chinese tracks, comparing against baseline prompting methods, and discuss the feasibility and limitations of lightweight agentic systems for clinical QA evaluation.

Keywords: evaluation, clinical applications, multimodal, multilingual, agentic AI

1. Background & Introduction

Digital messaging platforms have become an increasingly important component of modern healthcare delivery, enabling remote patient-provider communication and improving access to care while reducing logistical and financial barriers (Shickel et al., 2018; Topol, 2019). In visually driven medical specialties such as dermatology and wound care, patient queries frequently include images alongside free-text descriptions of symptoms, disease progression, or prior treatments (Yim et al., 2024c, 2025b). Automatically generating draft responses to such multimodal queries has thus become a promising approach for alleviating clinician workload and supporting scalable telemedicine.

Recent advances in large language models (LLMs) and vision-language models (VLMs) have substantially improved the fluency and expressiveness of generated medical text, as well as the ability to reason over visual inputs (Li et al., 2023; Singhal et al., 2023). However, reliably *evaluating* the quality of these free-text, multimodal responses remains an open research challenge. Unlike closed-form question answering (QA) or classification tasks, medical free-text responses may admit multiple valid answers reflecting different clinical opinions, levels of detail, or stylistic preferences. As a result, traditional automatic evaluation metrics such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) often correlate poorly with expert judgment in clinical settings (Peng et al., 2017; Yim et al., 2025a). This problem is further compounded in multimodal scenarios, where the correctness of a response depends on appropriate interpretation of clinical images in conjunction with textual context.

A growing body of work has emphasized the need for specialized evaluation frameworks that better align with human clinical assessments. Prior efforts have explored learned evaluation models, rubric-based scoring, and LLM-as-a-judge paradigms to assess factuality, completeness, and clinical relevance of generated outputs (Liu et al., 2023; Fu et al., 2023; Xu et al., 2025a). While LLM-as-a-judge approaches have demonstrated strong empirical performance across many text generation tasks, they typically rely on direct generative scoring in a single prompt-driven inference step. Such setups offer limited transparency into the reasoning process underlying a score and provide little structured evidence for downstream auditing or calibration. Moreover, scores may also be overly influenced by surface-level properties such as verbosity or stylistic alignment, and judgments may vary with prompt phrasing or contextual framing rather than underlying clinical correctness.

In parallel, several shared tasks have advanced multimodal medical QA by focusing on answer generation in dermatology and wound care, including DermaVQA and WoundcareVQA (Yim et al., 2024c, 2025b). While these benchmarks have driven progress in multimodal reasoning, previous tasks have primarily focused on generation perfor-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 141–152
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

mance and provide limited insight into the broader problem of response evaluation.

The MEDIQA-EVAL shared task (Ben Abacha and Yim, 2026), organized as part of the LREC 2026 ClinicalNLP Workshop, addresses this gap by shifting the focus from answer generation to evaluation of system-generated free-text responses. In this task, participants are given patient questions paired with one or more clinical images and are asked to assign quality scores to candidate answers produced by external systems. The task is both multimodal and multilingual, encompassing English and Chinese data, and evaluates responses along multiple clinically motivated dimensions, including overall quality, factual accuracy, completeness, relevance, and writing style. By grounding evaluation in expert human ratings, MEDIQA-EVAL provides a rigorous testbed for developing and analyzing automated evaluation methodologies in realistic settings.

In this paper, we describe Team BDI's submission to the MEDIQA-EVAL shared task. Our approach departs from the dominant paradigm of fine-tuning task-specific models for evaluation. Instead, we propose a modular, agent-based framework that treats evaluation as evidence-guided orchestration. The system is built around a ReAct-style agent (Yao et al., 2023) using an out-of-the-box LLM that integrates structured reasoning guided by task-specific prompts, multimodal retrieval of similar clinical encounters, auxiliary machine learning-based scoring signals, VLM generated references, and optional image augmentation tools. This design reflects recent trends toward agentic and compositional systems in applied natural language processing and explores their viability as lightweight and flexible alternatives for multimodal clinical QA evaluation.

Our contributions are three-fold:

- We introduce a ReAct-style, training-free evaluation framework for multimodal clinical QA that decomposes scoring into explicit stages of structured rubric-based reasoning, evidence retrieval, and integrative decision-making, rather than relying on monolithic end-to-end grading.
- We design an interpretable scoring mechanism that augments LLM-based reasoning with structured features and regression-based numeric priors, enabling more transparent, auditable, and potentially calibratable judgments compared to single-step LLM-as-a-judge approaches.
- We provide an empirical analysis on the multilingual MEDIQA-EVAL benchmark, showing that inference-time modular orchestration offers a viable alternative for multimodal clinical QA evaluation across English and Chinese settings that does not require any additional pre-training or fine-tuning of underlying foundation models.

2. Related Works

2.1. Medical Question Answering

Medical QA has evolved from traditional text-only retrieval or multiple-choice formats to increasingly complex multimodal tasks that integrate both visual and textual information (Lin et al., 2023). Early biomedical QA benchmarks focused on answering factual questions over text corpora or structured clinical data (Ben Abacha et al., 2018; Nentidis et al., 2024). However, recent work has expanded QA to incorporate multimodal reasoning, where systems must jointly interpret clinical images and accompanying questions to produce accurate answers (Demirhan and Zadrozny, 2024; Zhang et al., 2025). For example, large-scale medical visual QA efforts have targeted generation of natural language responses grounded in radiological, pathological, or endoscopic images paired with clinical queries (Zhang et al., 2024; Gautam et al., 2025).

2.2. Free-Text Generation and Evaluation

Free-text generation in medical QA introduces unique evaluation challenges due to variability in valid clinical responses and nuanced medical content. Several shared tasks have focused on multimodal free-text generation conditioned on clinical images and questions (Yim et al., 2024b; Xu et al., 2024). Similarly, the MEDIQA-WV and MEDIQA-MAGIC tasks have emphasized wound care and dermoscopy QA, requiring systems to produce fluent, clinically appropriate text accompanied by structured output where applicable (Yim et al., 2024a, 2025c).

Despite progress in generation, evaluation remains an active area of research. Generic automatic metrics often correlate poorly with clinical relevance and accuracy, leading to work on specialized evaluators and metrics (Van Veen et al., 2024; Ostmeier et al., 2024; Zhao et al., 2024; Xu et al., 2025b; Delbrouck et al., 2025). This underscores the need for evaluation benchmarks like MEDIQA-EVAL that directly target free-text quality in clinical QA.

2.3. Dermatology and Wound Care

Recent work has also explored foundation models and specialized datasets tailored to dermatology and wound care domains. In dermatology, the MM-Skin dataset enhances VLM capability by deriving large-scale image-text pairs from textbooks and

clinical sources, enabling richer skin disease representation and QA performance (Zeng et al., 2025). Other dermatology visual QA efforts such as DermaVQA provide multilingual benchmarks for skin condition reasoning (Yim et al., 2024c). In wound care, the WoundcareVQA dataset supports multimodal QA with structured wound attributes and clinical queries (Yim et al., 2025b), and recent retrieval augmented generation methods have been shown to be effective in the MEDIQA-WV shared task framework (Karim and Uzuner, 2025).

3. MEDIQA-EVAL Shared Task

The MEDIQA-EVAL shared task evaluates automatic methods for scoring free-text, multimodal clinical QA responses. Given a patient question paired with one or more clinical images and a set of candidate system-generated answers, participating systems assign numerical quality scores intended to approximate expert human judgements. The task is multilingual (English and Chinese) and focuses on visually driven clinical domains, specifically dermatology and wound care.

The benchmark builds on two previously released multimodal datasets: *DermaVQA* (referred to as `iiyi`), a dermatology visual question answering dataset, and *WoundcareVQA*, which focuses on wound care encounters. Both datasets consist of online patient queries accompanied by clinical images and clinician-authored responses. For the shared task, each encounter is paired with three candidate system outputs to be evaluated. A summary of the dataset composition used in the challenge is shown in Table 1.

Dataset	Split	#Encounters	#Systems	#Human
WoundcareVQA	Dev	105	315	210
	Test	93	279	279
DermaVQA (iiyi)	Dev	56	168	417
	Test	100	300	926

Table 1: Summary of datasets used in MEDIQA-EVAL. Each encounter is associated with three system-generated candidate responses and one or more human responses. Both English (EN) and Chinese (ZH) versions are provided for both development and testing splits.

Human expert annotations serve as the reference standard for evaluation. For English data, candidate responses are rated along multiple clinically motivated axes:

- **Completeness**: extent to which the response adequately addresses all clinically relevant aspects of the query.
- **Factual Accuracy**: clinical correctness of the medical information provided.
- **Relevance**: degree of alignment between the response and the patient's question.
- **Writing Style**: clarity, coherence, and appropriateness of the language used.
- **Overall Quality**: holistic assessment of the response's clinical usefulness and reliability.

Each of these five metrics is discretized to $\{0, 0.5, 1\}$, corresponding to poor, partial, or strong performance. In addition, the English data includes a binary **disagreement flag** indicating whether the evaluator *disagrees* with the overall clinical judgment or message conveyed in the response. Chinese samples are evaluated using a reduced set of criteria focusing on only factual consistency with human responses and writing quality. These dimensions are likewise discretized to $\{0, 0.5, 1\}$ and do not include a disagreement flag.

The development data was provided with multiple expert judgements per candidate. However, each candidate in the test data was independently annotated by two experts, with a conflict resolution process applied by the task organizers for cases of disagreement to produce the final reference scores used for evaluation.

Submitted automatic scoring systems are evaluated by measuring their correlation with human ratings. For each metric, Kendall's τ, Pearson's r, and Spearman's ρ are computed between system-generated scores and human judgements. These correlation coefficients are averaged to produce a metric-level score. Final leaderboard rankings are determined by averaging metric-level scores across all metrics within each language, yielding separate aggregate scores for English and Chinese. This evaluation protocol emphasizes relative ranking and calibration of candidate responses rather than absolute score prediction, encouraging systems to capture clinically meaningful distinctions among candidate answers.

4. Methods

4.1. Task Formulation

We formulate MEDIQA-EVAL as a multimodal response scoring problem. Let an encounter be defined as:

$$e = \langle q, I, M \rangle$$

where q denotes a patient query expressed as free text, $I = \{i_1, \ldots, i_{|I|}\}$ is a set of associated clinical images, and M represents optional structured metadata describing the encounter (*e.g.*, anatomical location or wound type). For each encounter e, a fixed set of candidate system responses $C = \{c_1, c_2, \ldots, c_K\}$ is provided, where

each c_k is a free-text answer generated by an external system.

The objective of the task is to assign numerical quality scores to each candidate c_k along a predefined set of evaluation dimensions. Let $\mathcal{D}$ denote the set of evaluation metrics, which differs by language (*i.e.*, six metrics for English and two for Chinese). The desired output of an automatic evaluation system is a function:

$$f : (e, c_k, d) \rightarrow \hat{y}_{k,d}$$

where $\hat{y}_{k,d} \in \mathbb{R}$ is the predicted score for candidate c_k on metric $d \in \mathcal{D}$. These predictions are evaluated by their rank and linear correlation with expert human judgements rather than absolute error, emphasizing relative ordering and calibration across candidates.

4.2. Evaluation Agent Overview

To address this task, we implement an *agent-based evaluation framework* that treats response scoring as a structured reasoning and orchestration problem rather than a purely parametric prediction task. At a high level, the agent operates as a policy:

$$\pi_\theta : (e, C) \rightarrow \{\hat{Y}_1, \ldots, \hat{Y}_K\}$$

where $\hat{Y}_k = \{\hat{y}_{k,d} \mid d \in \mathcal{D}\}$ denotes the vector of predicted metric scores for candidate c_k. The policy π_θ is instantiated by `gpt-5-2025-08-07` out-of-the-box, conditioned on a system prompt and augmented through external tools (Figure 1).

Following the ReAct paradigm, the agent interleaves natural language reasoning with tool invocations to gather auxiliary evidence before producing final scores. Rather than embedding all task-specific knowledge in model parameters, the agent dynamically queries specialized modules that expose structured signals relevant to clinical evaluation. This design emphasizes modularity and interpretability and avoids additional model fine-tuning.

Concretely, for each encounter e, the agent performs an iterative reasoning process: $s_{t+1} = \text{LLM}(s_t, o_t)$, where s_t denotes the agent's internal reasoning state at step t, and o_t represents observations returned by external tools. The process terminates when the agent emits a complete set of metric predictions for all candidates in C.

4.3. Model Context Protocol Tools

The agent interacts with external components through the Model Context Protocol (MCP) (Hou et al., 2025), which provides a standardized interface for invoking tools implemented as independent services. Each tool exposes a well-defined input-output contract and can be queried conditionally based on the agent's reasoning state. In our system, we employ several MCP servers corresponding to complementary sources of evidence: feature-based encounter retrieval, auxiliary machine learning-based scoring, VLM generated references, and image augmentation skills.

4.3.1. Feature-based Encounter Retrieval

Our retrieval subsystem returns a small set of previously observed encounters:

$$R(e) = \{e'_1, \ldots, e'_N\}$$

that are most relevant to an input encounter $e = \langle q, I, M \rangle$. Retrieval is implemented as a two-stage pipeline: (i) *metadata-based filtering* to produce a reduced candidate pool, followed by (ii) *embedding-based ranking* over text and images, with final fusion via reciprocal rank fusion (RRF) (Cormack et al., 2009).

Metadata Features. To enrich the datasets' sparse structured metadata, we first generated additional metadata annotations using `gpt-5.2-2025-12-11`. Specifically, for each encounter, we prompted the LLM to infer structured attributes from the query text and images, such as:

- **`query_type`**: coarse intent of the question (*e.g.*, diagnosis, management, urgency).
- **`anatomic_locations`**: normalized anatomical site(s) mentioned (list-valued).
- **`suspected_diagnosis`**: short phrase describing the likely diagnosis (scalar).
- **`contains_identifiable_feature`**: binary indicator whether the image shows an identifiable person/feature.

For metadata retrieval matching, we score store entries by counting exact matches for scalar fields and Jaccard similarity for list fields. Let M_{in} be the input metadata and M_{s} a store entry's metadata. We define the scalar-match component:

$$S_{\text{scalar}}(M_{\text{in}}, M_{\text{s}}) = \sum_{k \in \mathcal{K}_{\text{scalar}}} \mathbf{1}\big[M_{\text{in}}(k) = M_{\text{s}}(k)\big]$$

and the list-field component:

$$S_{\text{list}}(M_{\text{in}}, M_{\text{s}}) = \sum_{k \in \mathcal{K}_{\text{list}}} \frac{|A_k \cap B_k|}{|A_k \cup B_k|}$$

where A_k and B_k are the normalized sets for field k in the input and store entry respectively. The overall metadata score is the sum:

$$S_{\text{meta}}(M_{\text{in}}, M_{\text{s}}) = S_{\text{scalar}}(M_{\text{in}}, M_{\text{s}}) + S_{\text{list}}(M_{\text{in}}, M_{\text{s}})$$

We select the top $N_{\text{meta}} = 10$ store entries ranked by S_{meta} (ties broken deterministically) to form the candidate pool for the next stage.

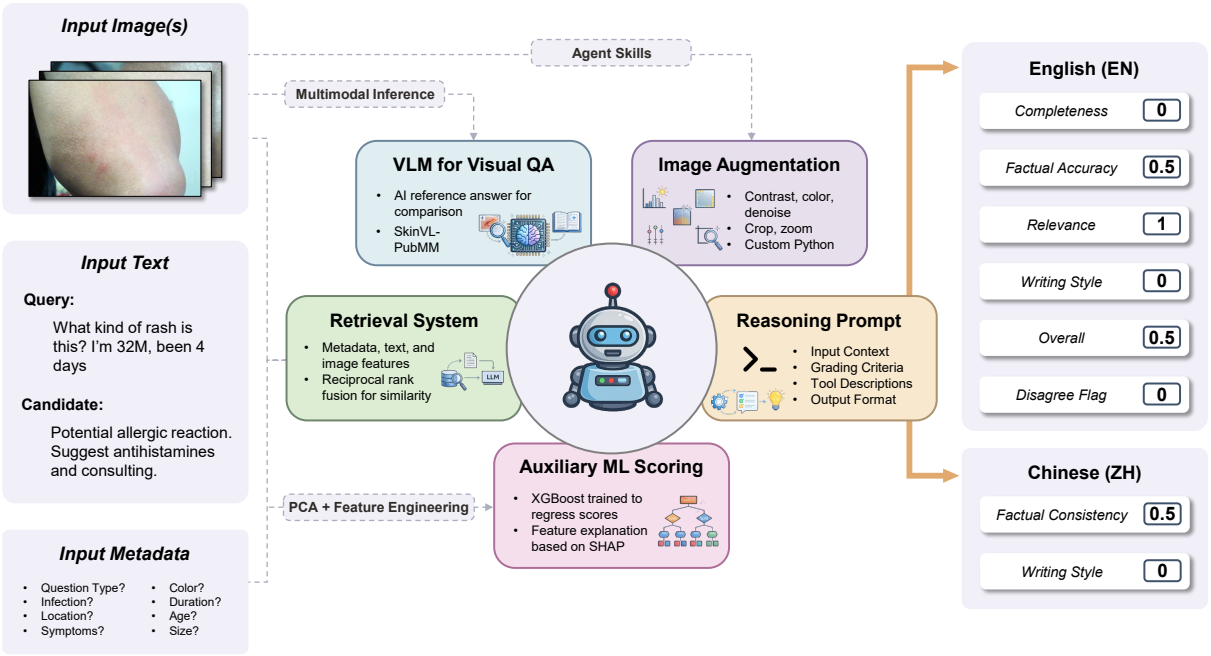


Figure 1: Schematic of the evaluation agent showing various tool modules. A representative English encounter is shown. For Chinese encounters, the input query and candidate would be translated, and the output would only consist of two metrics.

Textual Features. For textual similarity, we embed full queries and candidate answer text using `paraphrase-multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2019). In addition to dense embeddings, we perform light clinical entity extraction using *regex* and keyword lists (*e.g.*, age, duration, and anatomical terms) to provide interpretable overlap features. Given a query embedding $\mathbf{q} \in \mathbb{R}^d$ and a store text embedding $\mathbf{t} \in \mathbb{R}^d$, we compute cosine similarity:

$$\text{sim}_{\text{text}}(\mathbf{q}, \mathbf{t}) = \frac{\mathbf{q} \cdot \mathbf{t}}{\|\mathbf{q}\|\|\mathbf{t}\|}$$

Image Features. For visual similarity we include two complementary signals:

- *MedImageInsight* embeddings (Codella et al., 2024) extracted via an internal endpoint. These capture clinically relevant visual representations and serve as the primary medical image embedding source.
- *ViT CLIP* embeddings (Radford et al., 2021) to provide a general-purpose visual-text alignment signal.

We additionally compute descriptive image statistics (*e.g.*, RGB mean/std and HSV means) and texture descriptors (*e.g.*, patch-level variance and simple saliency patch counts). For multi-image encounters we represent the visual evidence as the set of per-image embeddings $\{\mathbf{i}_1, \ldots, \mathbf{i}_m\}$. The image-to-image similarity between input images and a store encounter is computed by taking the maximum pairwise cosine similarity across all image pairs:

$$\text{sim}_{\text{image}}(I, I_s) = \max_{u \in I} \max_{v \in I_s} \frac{\mathbf{i}_u \cdot \mathbf{j}_v}{\|\mathbf{i}_u\|\|\mathbf{j}_v\|},$$

which naturally handles differing image counts per encounter by focusing on the best-matching pair.

Ranking and Fusion. After the metadata filtering stage produces candidate encounters $\mathcal{C}_{\text{meta}}$, we produce two ranked lists over $\mathcal{C}_{\text{meta}}$: a text-based ranking using sim_{text} and an image-based ranking using $\text{sim}_{\text{image}}$. Each ranking is an ordered list with ranks $r_{\text{text}}(e')$ and $r_{\text{image}}(e')$ for $e' \in \mathcal{C}_{\text{meta}}$. We fuse the two rankings using RRF:

$$\text{score}_{\text{RRF}}(e') = \sum_{m \in \{\text{text,image}\}} \frac{1}{k + r_m(e')},$$

where k is an RRF parameter (we use $k = 60$ in our implementation). The fused ordering is obtained by sorting $\text{score}_{\text{RRF}}$ in descending order, and the top N fused encounters are returned to the agent as exemplars $R(e)$. In practice, we return the top 3 fused exemplars.

4.3.2. Auxiliary Machine Learning-based Scoring

To supply the agent with compact, explainable numeric reference signals, we trained auxiliary

gradient-boosted decision tree models (XGBoost (Chen and Guestrin, 2016)) that predict metric-specific scores. These models operate on a dense, precomputed feature representation $\phi(e, c_k) \in \mathbb{R}^M$ constructed from the text, image, and metadata pipelines described above. For each evaluation metric d (*e.g.*, `completeness`, `factual-accuracy`, and `overall`), we fit a regression model:

$$g_d : \mathbb{R}^M \to \mathbb{R}, \qquad \tilde{y}_{k,d} = g_d\big(\phi(e, c_k)\big)$$

where $\tilde{y}_{k,d}$ is the predicted score used as an auxiliary signal by the agent.

Feature Engineering. The feature matrix is created in two stages. First, we assemble a base row for every (encounter, candidate) pair containing:

- **Text features:** query and candidate dense embeddings reduced with PCA to $d_{\text{PCA}} = 32$ dimensions, discrete query type, entity counts, candidate length, and punctuation statistics.
- **Image features:** mean-pooled per-image *MedImageInsight* embeddings reduced with PCA to $d_{\text{PCA}} = 32$, saliency/patch statistics, and image tag one-hots.
- **Metadata features:** one-hot / count encodings for anatomic locations, wound type, and author/rater statistics.

Second, we augment the base matrix with interpretable and interaction features, such as Flesch reading ease (English), average tokens-per-sentence, counts of medical terms, action recommendation, harmful-recommendation heuristics, cross-modal cosine similarities (between query-image), mean/max textual cosine similarities (between candidate-gold), and triangle-style features (cosine similarity of `query_candidate` minus cosine similarity of `image_candidate`).

Training Objective and Explainability. For each metric d on the validation set, we use the mean human rating (aggregated across raters for that candidate) as the regression target. Models were trained using XGBoost with squared error objective. We train separate models per metric and validate using held-out validation folds derived from the provided development split.

For each prediction $\tilde{y}_{k,d}$, we also compute SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) values to extract the top-K explainable features (excluding PCA/embedding dimensions) that contributed most to the model output.

Agentic Model Usage. During per-encounter evaluation, the ReAct agent invokes the XGBoost MCP to retrieve $\{\tilde{y}_{k,d}\}_d$ and the associated top-5 SHAP features for each candidate in natural language form to guide downstream reasoning, which was previously shown to aid in feature engineering for tabular clinical prediction tasks (Zhang et al., 2026). The agent conditions on these signals; however, it is free to accept, down-weight, or contradict the auxiliary predictions. We designed this interaction to provide a structured, interpretable corrective signal (in the spirit of explainability-driven model steering) without hard-wiring the final outputs to the auxiliary regressor.

4.3.3. VLM-based Visual QA Reference

We also incorporate a VLM-based visual QA reference. Given an encounter $e = \langle q, I, M \rangle$, we generate a VLM-based answer using a pre-trained multimodal foundation model specialized for dermatological reasoning (`zwq803/SkinVL-PubMM` (Zeng et al., 2025)). The model takes the query text and associated clinical images as input and produces a free-text response intended to approximate a clinically appropriate answer.

The generated response is not treated as ground truth. Instead, it serves as an additional reference signal that the evaluation agent can compare against candidate responses. The agent may assess semantic similarity between a candidate c_k and the generated answer, or use discrepancies between them as evidence when evaluating factual accuracy or completeness.

4.3.4. Image Augmentation Skills

Agent skills (Anthropic, 2025) expose a lightweight image augmentation toolkit that the agent may invoke when visual evidence is insufficient (*e.g.*, low contrast, poor lighting, or small lesion sizes) or when a localized patch requires closer inspection. Augmentations are treated as deterministic image-to-image transformations $i'_j = T(i_j; \psi)$, where i_j is an input image, T is an augmentation operator selected by the agent, and ψ denotes operator parameters (*e.g.*, Contrast Limited Adaptive Histogram Equalization (CLAHE (Pizer et al., 1990)) clip limit or gamma). Augmented images are returned to the agent for inspection and reasoning.

Invocation and Augmentations. The agent supplies a short, self-contained Python code snippet that reads the input image, applies image-processing operations (using `Pillow`, `cv2`, and/or `numpy`), and writes the result. The snippet runs inside a sandboxed subprocess with no external network or arbitrary filesystem access. The agent then loads the augmented image and continues

reasoning. This design keeps augmentations explicit, auditable, and reversible. We implemented and recommended a small set of clinically useful operations that the agent may combine sequentially:

- **Contrast enhancement / CLAHE.** Apply CLAHE to local image tiles to improve visibility of subtle texture or color differences.
- **Edge / border emphasis.** Sobel or Laplacian filters (optionally blended with the original image) to highlight lesion margins.
- **Color normalization / channel emphasis.** Scale or reweight RGB/Hue channels to emphasize erythema or pigmentation (*e.g.*, amplify red channel or threshold HSV ranges).
- **Denoising.** Apply bilateral or non-local means denoising prior to contrast steps to avoid amplifying noise.
- **Crop / zoom to region-of-interest (ROI).** Crop and resize to create a close-up patch for finer inspection.

The agent decides to invoke augmentation based on internal signals and metadata: *e.g.*, the agent inspects `contains_identifiable_feature` or computed image statistics (low contrast, low mean brightness, or low saliency counts) and may request an augmentation if these indicators suggest improvement is possible. The agent is instructed to keep augmentations *moderate* to avoid introducing artifacts that could mislead clinical interpretation. In practice, these augmentations modestly improve the agent's ability to (i) detect lesion borders and exudate, (ii) disambiguate pigmentation vs. shadowing, and (iii) increase the signal-to-noise ratio of patch-level texture statistics. Because augmentations are fast (simple `OpenCV`/`Pillow` operations) and executed only on demand, they provide a computationally cheap way to increase visual interpretability without retraining visual encoders.

5. Results

5.1. English Subtask

Table 2 summarizes the English leaderboard results. Our system ranked 4th out of 5 teams on the aggregate mean score across all metrics. Only one participating team surpassed the organizer-provided baseline on the mean-of-all-metrics score. The strongest systems in this track adopted supervised fine-tuning strategies using the development data. In contrast, our submission relied entirely on prompting and modular tool orchestration without additional model training.

Although our aggregate ranking was lower, several metric-level patterns are worth highlighting. Our system performed competitively on the `writing-style` metric and achieved relatively strong results on the `overall` metric compared to other non-fine-tuned approaches. This suggests that the agent's structured reasoning and auxiliary signals were effective at capturing holistic quality and stylistic alignment.

However, performance was substantially lower on the `completeness` metric. We observed that the agent tended to assign lower completeness scores to concise human-generated answers, likely reflecting a verbosity bias inherent to LLMs, especially if the agent heavily conditioned its scores on visual QA references generated by LLMs. Because `completeness` contributes directly to the aggregated mean with equal weight to other metrics, it had a noticeable impact on the overall ranking.

To better understand the contribution of individual modules, we conducted iterative internal experiments during development. Starting from a baseline prompting setup (GPT-4o; for "All Metrics", EN: 0.186, ZH: 0.125), we incrementally added (i) retrieval, (ii) auxiliary XGBoost scoring, (iii) LLM generated references, and (iv) image augmentation tools. Each addition produced measurable improvements in correlation on our held-out set of the development split, indicating that the MCP-based modular components contribute positively even without explicit fine-tuning.

Rank	Team	All Metrics	Overall Metric
1	SUAT-BMI	0.482	0.545
2	SloCal-Net	0.416	0.466
3	MedAware	0.391	0.526
4	**BDI**	**0.369**	**0.439**
5	hgkai26	0.213	0.244
–	MEDIQA (Baseline)	0.451	0.555

Table 2: MEDIQA-EVAL leaderboard results for the English subtask.

5.2. Chinese Subtask

Table 3 presents the Chinese results. Our system ranked 2nd out of 3 teams. Unlike the English track, performance differences among teams were comparatively smaller. We observed particularly strong performance on the `factual-consistency-wgold` metric, indicating that the retrieval-augmented reasoning and candidate-gold similarity features transfer well to Chinese.

In contrast, the system underperformed on the `writing-style` metric. One plausible explanation is that stylistic expectations for Chinese online medical responses differ from the model's dominant training distribution. Since our system relies

on inference-time prompting rather than language-specific fine-tuning, stylistic calibration may be less precise in Chinese. Nevertheless, the narrow performance gap suggests that prompting-based approaches with multilingual models can remain competitive without language-specific retraining.

Overall, the Chinese results highlight a potential advantage of modular, inference-driven systems: a single architecture can generalize across languages with minimal additional engineering effort, whereas fine-tuning approaches may require separate models or additional multilingual supervision.

Rank	Team	All Metrics
1	SloCal-Net	0.294
2	**BDI**	**0.269**
3	MedAware	0.259
–	MEDIQA (Baseline)	0.277

Table 3: MEDIQA-EVAL leaderboard results for the Chinese subtask.

6. Discussion and Conclusion

We presented a modular, agent-based framework for multimodal free-text evaluation in the MEDIQA-EVAL shared task. Importantly, all components operate without additional model pre-training or fine-tuning. Task specificity is achieved through prompt design, tool orchestration, and feature engineering. This architecture enables rapid experimentation and highlights the feasibility of compositional, agentic systems for complex clinical evaluation tasks.

From a systems perspective, several design choices proved beneficial. All embeddings, metadata annotations, and document records are stored in offline indexed stores, with embedding shards and encounter-level mappings to enable efficient lookup during inference. This design minimizes redundant computation and allows scalable retrieval across encounters.

The two-stage retrieval strategy – metadata filtering followed by embedding-based ranking with RRF – reduces unnecessary cross-domain comparisons while leveraging LLM-augmented metadata to capture clinically salient attributes absent from the original dataset-supplied fields. By fusing text and image rankings, the system remains robust to modality imbalance (*e.g.*, encounters with multiple images or limited textual detail). Retrieved exemplars provide contextual human responses and historically scored candidates, supporting retrieval-augmented reasoning and structured evaluation.

The auxiliary XGBoost module complements reasoning by providing structured numeric priors and feature-level explanations via SHAP. Rather than constraining outputs, these signals act as interpretable guidance, allowing the agent to integrate model-based predictions with broader contextual reasoning. This hybrid design reflects a middle ground between fully parametric fine-tuning and purely prompt-based scoring.

Future Directions. There are several avenues for improvement. First, stronger domain-specific visual encoders, particularly dermatology- or wound-care-specialized foundation models, may yield improved image embeddings and cross-modal similarity features. Our current approach relies on general-purpose or proprietary embeddings. Domain-tuned representations could better capture subtle clinical cues.

Second, additional prompt calibration may better align LLM grading behavior with human annotation guidelines, particularly for metrics such as completeness and writing style. Explicit rubric reminders, metric-specific reasoning steps, or stylistic exemplars may improve alignment without requiring parameter updates.

Third, the evaluation problem itself could be decomposed into more constrained sub-questions. For example, instead of directly predicting a scalar completeness score, the system could assess whether key clinical components (diagnosis, management advice, risk discussion) are present and then aggregate those assessments.

Our current framework relies on a small, fixed number of retrieved exemplars and predefined fusion strategies, but we did not exhaustively explore the sensitivity of performance to these choices. Future work should conduct a rigorous ablation study examining the number of retrieved samples, metadata filtering thresholds, and text-image fusion weights, in order to better understand their contribution to final correlation scores.

Finally, synthetic data generation or expanded human annotation could help stabilize evaluation models. The development split, which the retrieval system heavily depends on, is limited in size and exhibits annotation variability. Generating additional high-quality examples, either via expert labeling or carefully controlled LLM simulation, may reduce noise sensitivity for both fine-tuning and feature-based approaches.

While our system did not achieve top leaderboard performances, the results demonstrate that modular, inference-driven agents can still be a viable method for multilingual settings with minimal retraining overhead. As multimodal evaluation benchmarks continue to evolve, compositional agent frameworks offer a flexible and extensible alternative to fine-tuned models.

7. Limitations

Dependence on noisy development data. Our retrieval-based strategy relies heavily on the quality of the development split. Although the organizers noted that the development annotations contain noise, our system retrieves at most ten metadata-matched encounters and ultimately conditions on only the top few exemplars during inference. Consequently, if the retrieved encounters contain inconsistent or noisy human ratings, this noise directly propagates into the agent's reasoning. In contrast, fine-tuning approaches implicitly learn aggregate patterns across the entire development split, potentially smoothing over local inconsistencies. Retrieval-based systems are therefore particularly sensitive to annotation quality. A possible improvement would be to curate a smaller, high-quality subset of development encounters with adjudicated ratings. However, given the inherent variability in clinician preferences, such curation may still not perfectly align with the hidden test annotations.

Inference latency and engineering overhead. The architecture depends on multiple model calls interleaved with local tool execution. Although token usage is moderate and does not require GPU-based model training, wall-clock inference time can be substantial due to sequential LLM inferences and MCP tool invocations. This may limit scalability in real-time deployment scenarios. Engineering optimizations (parallel tool execution, caching, batching, or leveraging smaller models such as `gpt-5-mini`) could substantially reduce latency. Preliminary experiments suggest that smaller frontier models can still reason effectively with tool calls, though with modest performance trade-offs.

Multilingual stylistic calibration. While the system generalizes well across languages without retraining, stylistic calibration – particularly for the Chinese writing style metric – remains imperfect. Available LLMs are predominantly trained on English-centric corpora, and stylistic expectations for online clinical communication may vary by language and cultural context. Prompt refinement or language-specific stylistic exemplars may improve alignment.

Subjectivity of clinical evaluation. More fundamentally, free-text clinical evaluation is inherently subjective. Clinician ratings depend on training background, regional practice patterns, and individual communication preferences. There is no universally standardized rubric that fully captures "correctness" or "quality" across all practitioners. As a result, even well-designed automatic systems may struggle to achieve high correlation across diverse annotator populations. This challenge extends beyond MEDIQA-EVAL and reflects a broader open research problem in evaluating generative medical AI systems.

Scope of visual representations. Our visual representations rely on general-purpose or externally trained embeddings rather than dermatology- or wound-care-specific vision models trained directly on the task distribution. Although we incorporate *MedImageInsight* and *CLIP* embeddings, further domain adaptation or use of highly specialized medical vision encoders could potentially improve cross-modal alignment.

Ultimately, while the proposed agent-based framework provides a flexible and extensible alternative to fine-tuning approaches, its performance remains sensitive to annotation quality, contextual limitations, and the intrinsic subjectivity of clinical free-text evaluation.

8. Acknowledgments

This project was funded by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with the UK Health Security Agency (UKHSA) (NIHR207397) and supported by the NIHR Biomedical Research Centre, Oxford. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health and Social Care, or UKHSA. JX gratefully acknowledges joint support from Canadian Institutes of Health Research (CIHR) Project ID 202410BCB-535721-77482 (Bioinformatics and Computational Biology), Nuffield Department of Medicine (NDM), and Oxford University Press (OUP). ZZ acknowledges support from Nuffield Department of Population Health (NDPH) and OUP. AL is a recipient of the University of Oxford Croucher Scholarship. HC is supported by the Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (CIFMS) Grant Number 2024-I2M-2-001-1.

9. Bibliographical References

Anthropic. 2025. Agent skills. https://platform.claude.com/docs/en/agents-and-tools/agent-skills/overview. Claude API Docs. Accessed: 2026-02-11.

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2018. *Overview of the Medical Question Answering Task at TREC 2017 LiveQA*. National Institute of Standards and Technology (NIST).

Asma Ben Abacha and Wen-wai Yim. 2026. Overview of the mediqa-eval 2026 shared task on evaluation metrics in medical multimodal question answering. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. European Language Resource Association (ELRA).

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.

Noel Codella, Yu Gu, Shrey Jain, Ho Hin Lee, Asma Ben Abacha, Alberto Santamaria-Pang, Will Guyman, Natieek Sangani, Sheng Zhang, Hoifung Poon, Stephanie Hyland, Shruthi Bannur, Javier Alvarez-Valle, Xue Li, John Garett, Alan McMillan, Gaurav Rajguru, Madhu Maddi, Nilesh Vijayrania, Reehan Bhimai, Nick Mecklenburg, Rupal Jain, Daniel Holstein, Naveen Gaur, Vijay Aski, Jenq-Neng Hwang, Thomas Lin, Ivan Tarapov, Matthew P. Lungren, and Mu Wei. 2024. MedImageInsight: An open-source embedding model for general domain medical imaging. *arXiv*.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 758–759. Association for Computing Machinery.

Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, Eduardo Pontes Reis, Mohamed S Muneer, Maya Varma, and Curtis Langlotz. 2025. Automated structured radiology report generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26813–26829. Association for Computational Linguistics.

Hilmi Demirhan and Wlodek Zadrozny. 2024. Survey of multimodal medical question answering. *BioMedInformatics*, 4(1):50–74. Publisher: Multidisciplinary Digital Publishing Institute.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Sushant Gautam, Michael A. Riegler, and Pål Halvorsen. 2025. Kvasir-VQA-x1: A multimodal dataset for medical reasoning and robust MedVQA in gastrointestinal endoscopy.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. 2025. Model context protocol (mcp): Landscape, security threats, and future research directions.

A. H. M. Rezaul Karim and Ozlem Uzuner. 2025. Multimodal retrieval-augmented generation with large language models for medical VQA.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611.

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.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Salvador Lima-López, Eulàlia Farré-Maduell, Martin Krallinger, Natalia Loukachevitch, Vera Davydova, Elena Tutubalina, and Georgios Paliouras. 2024. Overview of BioASQ 2024: The twelfth BioASQ challenge on large-scale biomedical semantic indexing and question answering.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. 2024. Green: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 374–390. 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. Association for Computational Linguistics.

Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

S.M. Pizer, R.E. Johnston, J.P. Ericksen, B.C. Yankaskas, and K.E. Muller. 1990. Contrast-limited adaptive histogram equalization: speed and effectiveness. In *[1990] Proceedings of the First Conference on Visualization in Biomedical Computing*, pages 337–345.

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.

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*. Association for Computational Linguistics.

Benjamin Shickel, Patrick James 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.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

Eric J. Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56. Publisher: Nature Publishing Group.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142. Publisher: Nature Publishing Group.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98.

Justin Xu, Yiming Li, Zizheng Zhang, Augustine Yui Hei Luk, Mayank Jobanputra, Samarth Oza, Ashley Murray, Meghana Reddy Kasula, Andrew Parker, and David W Eyre. 2025a. Tree-of-quote prompting improves factuality and attribution in multi-hop and medical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5605–5622, Suzhou, China. Association for Computational Linguistics.

Justin Xu, Xi Zhang, Javid Abderezaei, Julie Bauml, Roger Boodoo, Fatemeh Haghighi, Ali Ganjizadeh, Eric Brattain, Dave Van Veen, Zaiqiao Meng, David W Eyre, and Jean-Benoit Delbrouck. 2025b. RadEval: A framework for radiology text evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 546–557, Suzhou, China. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, and Fei Xia. 2024a. Overview of the MEDIQA-MAGIC task at ImageCLEF 2024: Multimodal and generative TelemedICine in dermatology.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2025a. Morqa: Benchmarking evaluation metrics for medical open-ended question answering.

Wen-wai Yim, Asma Ben Abacha, Robert Doerning, Chia-Yu Chen, Jiaying Xu, Anita Subbarao, Zixuan Yu, Fei Xia, M. Kennedy Hall, and Meliha Yetisgen. 2025b. Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care. *Journal of Biomedical Informatics*, 170:104888.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen, and Martin Krallinger. 2024b. Overview of the MEDIQA-m3g 2024 shared task on multilingual multimodal medical answer generation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 581–589. Association for Computational Linguistics.

Wen-wai Yim, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2025c. Overview of the MEDIQA-WV 2025 shared task on woundcare visual question answering. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 17–21. Association for Computational Linguistics.

Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2024c. DermaVQA: A Multilingual Visual Question Answering Dataset for Dermatology . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005. Springer Nature Switzerland.

Wenqi Zeng, Yuqi Sun, Chenxi Ma, Weimin Tan, and Bo Yan. 2025. MM-skin: Enhancing dermatology vision-language model with an image-text dataset derived from textbooks.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277. Publisher: Nature Publishing Group.

Zhilin Zhang, Jie Wang, Zhanghao Qin, Ruiqi Zhu, and Xiaoliang Gong. 2025. Efficient bilinear attention-based fusion for medical visual question answering.

Zizheng Zhang, Yiming Li, Justin Xu, Jinyu Wang, Rui Wang, Lei Song, Jiang Bian, David W Eyre, and Jingjing Fu. 2026. Medfeat: Model-aware and explainability-driven feature engineering with llms for clinical tabular prediction.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. RaTEScore: A metric for radiology report generation. Pages: 2024.06.24.24309405.

MasonNLP at MEDIQA-SYNUR 2026: Retrieval-Augmented Large Language Models for Schema-Constrained Clinical Information Extraction

A H M Rezaul Karim, Özlem Uzuner
George Mason University, VA, USA
akarim9@gmu.edu, ouzuner@gmu.edu

Abstract

Conversational nurse-patient transcripts contain actionable observations, but converting these transcripts into structured representations at scale remains challenging. Documentation burden is substantial, with prior studies showing clinicians spend large portions of their workday on documentation and related desk work rather than direct patient care. MEDIQA-SYNUR focuses on observation extraction from conversational nurse-patient transcripts, requiring systems to normalize these narratives into a predefined schema with value-type constraints. We propose a modular retrieval-augmented generation (RAG) pipeline that uses the training set as an exemplar corpus, combines schema-constrained prompting (full schema vs. pruned candidate schema), deterministic schema-based postprocessing, and a second-pass audit, with two LLM backbones: *Llama-4-Scout-17B-16E-Instruct* and *GPT-5.2* with corresponding embedding models for RAG. Our best configuration uses *GPT-5.2* with full schema, RAG, and a second-pass auditing, achieving 80.36% F_1 score. Overall, our results show that RAG consistently improves performance, while the optimal degree of schema constraint depends on the model, and second-pass auditing yields modest additional gains by correcting residual schema-adherence errors.

Keywords: Retrieval Augmented Generation (RAG), Clinical Information Extraction, LLM

1. Introduction

Clinical documentation is essential for continuity of care and quality reporting, yet it remains a major source of clinician workload and professional dissatisfaction (Muhiyaddin et al., 2022). A study in ambulatory practice has shown that physicians spend nearly half of their clinic day on electronic health record (EHR) and desk work (49.2%) and substantially less time in direct clinical face time (27.0%), with additional after-hours work of approximately 1-2 hours per night devoted mostly to EHR tasks (Sinsky et al., 2016). Similar EHR log-based analyses report that primary care clinicians spend 145.9 minutes/day actively using the EHR (Rotenstein et al., 2022). These high demands for documentation motivate methods that can automatically capture and convert clinical information into structured representations with minimal manual entry. As a step in this direction, particularly for workflows such as nursing assessments, clinical documentation is frequently transcribed (Mayer et al., 2022). Conversational nurse-patient transcripts are examples of such documentation.

Information extraction (IE) can convert transcripts into structured representations (Balasubramanian et al., 2025; Hu et al., 2026). In many clinical IE settings, the goal is to identify spans, assign concept labels, or extract relations from clinical text, often within relatively limited label spaces and without requiring the model to generate a fully normalized structured output. Strict requirements on output, such as conforming to a large predefined schema with consistent formatting and validity constraints (Karim and Uzuner, 2025; Builtjes et al., 2025) pose additional challenges for clinical IE.

MEDIQA-SYNUR task presents an observation extraction task that encompasses these challenges, with a specific focus on conversational nurse-patient transcripts. Given a transcript, MEDIQA-SYNUR requires automated systems to extract clinically salient observations normalized to a predefined schema with explicit value-type requirements (e.g., numeric vs. categorical values and enumerated option validity) (Michalopoulos et al., 2026). The *schema* for the task captures the complete set of target observation concepts together with their value types and any allowable categorical value sets. The SYNUR dataset accompanying this task provides an open-source synthetic corpus annotated by expert nurses with a large set of structured observations, enabling controlled evaluation of observation extraction from conversational nurse-patient transcripts (Corbeil et al., 2025).

Recent advances in instruction-following large language models (LLMs) have enabled prompt-based clinical IE for converting clinical narratives into structured representations (Agrawal et al., 2022; Rodrigues and Teixeira Lopes, 2025). Compared with task-specific supervised systems, LLMs offer a more flexible framework, and prior work has shown that, with carefully designed prompts, they can support few-shot extraction across clinical IE tasks (Agrawal et al., 2022; Rodrigues and Teixeira Lopes, 2025). However, reliability and schema adherence remain persistent concerns, particularly

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 153–162
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

as output spaces grow. LLMs tend to hallucinate concept names, violate type constraints, and produce inconsistent formatting when schema complexity increases (Karim and Uzuner, 2025; Builtjes et al., 2025). Evidence also suggests that open-weight and closed-weight models exhibit meaningfully different instruction-following behaviors and robustness profiles under constraint-heavy prompts (Builtjes et al., 2025). In this work, we explore open-weight and closed-weight LLMs in observation extraction by constraining their output through a schema that varies in its complexity. We refer to this as schema-constrained observation extraction.

We hypothesize that retrieval-augmented generation (RAG) (Lewis et al., 2020) can improve performance in observation extraction by conditioning output on retrieved exemplars (Shlyk et al., 2024; Lopez et al., 2025; Zhan et al., 2025; Liu et al., 2025). To test this hypothesis, we study two RAG (Lewis et al., 2020) approaches that use the training set as the retrieval corpus. For *Llama-4-Scout-17B-16E-Instruct* (Meta, 2025), we find that constraining the output space using a pruned candidate schema improves results. In contrast, with *GPT-5.2* (OpenAI, 2025), RAG yields the best performance with full schema, and a second-pass auditing provides additional gains.[1]

Our study makes the following contributions:

- The question of the interaction of RAG with the schema is in Clinical IE. Our work directly addresses this gap by studying how RAG interacts with different schema constraints across LLM backbones of differing parameter scales, providing a systematic analysis of this interaction in observation extraction.
- We show that a pruned candidate schema is beneficial for a smaller open-weight model but counterproductive for a larger model, suggesting that the formulation of the output constraints, namely how the fixed task schema is presented to the model during generation, should be model-aware, i.e., adapted to the behavior and capabilities of the underlying model.
- We present second-pass auditing that provides modest gains primarily by correcting schema-adherence and normalization errors, but does not substitute for RAG and schema constraint, indicating it is best used as a final refinement stage rather than a core component.

Overall, our findings provide practical evidence that schema-constrained observation extraction should be model-aware. Our results clarify when a pruned candidate schema is helpful versus when a full schema is preferable, and further show how RAG and second-pass auditing can be combined to improve robustness.

[1] Implementation details can be found here: https://github.com/AHMRezaul/MEDIQA-SYNUR-2026

2. Related Work

Clinical information extraction (IE) has been driven by tasks that standardize evaluation for concept extraction, assertion detection, and relations (Uzuner et al., 2011; Henry et al., 2020; Fu et al., 2020; Navarro et al., 2023). Benchmarks such as i2b2/VA and n2c2 established protocols for span extraction and normalization, but largely focus on relatively small label spaces (Uzuner et al., 2011; Henry et al., 2020; Mahajan et al., 2023). In contrast, MEDIQA-SYNUR focuses on a large, heterogeneous output space (193 typed observation concepts), which must be extracted under strict type and enumeration constraints, motivating methods that explicitly condition generation on schema.

Early studies of conversational nurse-patient transcripts coupled speech recognition with IE to produce structured handover documents, highlighting both feasibility and the difficulty of extracting structured representations from transcripts (Johnson et al., 2014a,b; Dawson et al., 2014). Subsequent shared evaluations released transcribed handover datasets and structured annotations, often using synthetic data to mitigate privacy barriers (Suominen et al., 2015b,a, 2016). Building on this line, Corbeil et al. (2025) introduced SYNUR to support systematic evaluation of structured output from conversational nurse-patient transcripts, leaving open how to enforce schema adherence reliably at scale.

Within LLM-based clinical IE, most prior work has emphasized direct prompting or few-shot extraction, often in settings where outputs are interpreted more flexibly and are not tightly constrained by a large predefined schema (Agrawal et al., 2022; Rodrigues and Teixeira Lopes, 2025). Work that explicitly examines schema adherence has shown that structured extraction becomes substantially more difficult as output spaces grow and validity constraints become stricter (Karim and Uzuner, 2025; Builtjes et al., 2025). Although retrieval-augmented generation has begun to show promise in clinical extraction by grounding predictions in relevant examples or evidence (Lewis et al., 2020; Shlyk et al., 2024; Lopez et al., 2025; Zhan et al., 2025; Liu et al., 2025), its interaction with schema-constrained generation remains underexplored. Likewise, limited prior work has compared how open-weight and closed-weight LLMs behave under such constraint-heavy extraction settings. Our work addresses these gaps in the context of schema-constrained observation extraction.

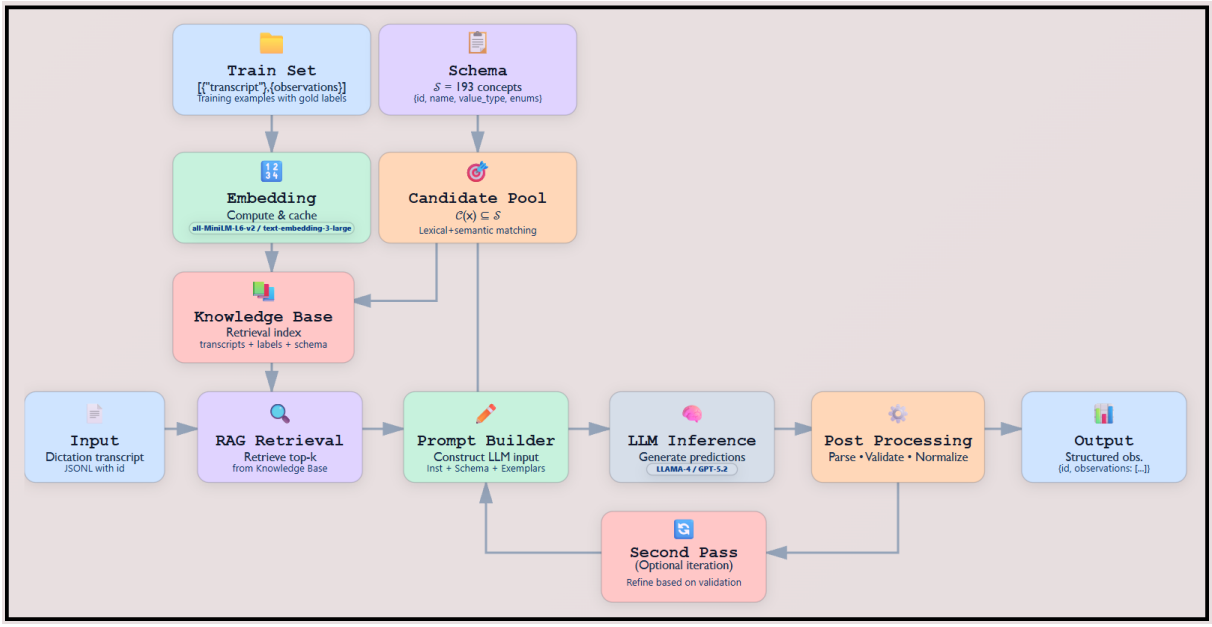


Figure 1: Retrieval-augmented, schema-constrained pipeline that retrieves training exemplars, conditions the LLM on schema (full or pruned candidate), and post-processes outputs with second-pass auditing.

3. Task Description

3.1. Problem Formulation

MEDIQA-SYNUR defines the task as an *observation extraction* from conversational nurse-patient transcripts. Given a transcript x, the goal is to identify clinically salient observations and normalize them to a predefined schema.

The schema is a set of $M = 193$ observation concepts $\mathcal{S} = \{c_1, \ldots, c_M\}$. Each concept c_m has an identifier id_m, a name N_m, a value type $\tau_m \in \{\textsc{single_select}, \textsc{multi_select}, \textsc{numeric}, \textsc{string}\}$, and, for categorical types, an allowed value set $\mathcal{V}_m$.

For an input x, a system outputs a list of extracted observation instances $\hat{O} = [o_1, \ldots, o_n]$, where each instance is an object from the schema $o_i = \{\hat{\mathrm{id}}_i, \hat{N}_i, \hat{\tau}_i, \hat{v}_i\}$. The evaluation measures the correctness of the predicted observations under this schema.

3.2. Dataset

The SYNUR dataset (SYnthetic NURsing) (Corbeil et al., 2025) is a JSONL file split into `train`, `dev`, and `test`. Table 1 contains the dataset statistics of each split. Each instance contains a unique case identifier `id` and a free-text `transcript`. It also contains ground truth `observations`, represented as a list of objects from the schema with values that may be categorical (single- or multi-select), numeric, or free-text, reflecting the heterogeneity of routine nursing documentation.

Split	No. of Inst.	Total Obs.	Unique Obs.
Train	122	1685	166
Dev	101	1315	170
Test	199	2552	164

Table 1: Dataset statistics per split with the number of instances, total number of observations, and the number of unique observations in each split.

Across all splits, conversational nurse-patient transcripts are moderately long, averaging 192 words, with lengths ranging from 59 to 343 words. In the train and development sets, each instance contains 13.45 observations on average, ranging from 6 to 34 observations per case, with no duplicated concept IDs within an instance. Label frequency follows a long-tailed distribution, while a small set of observations occur very frequently (e.g., *Cognitive status*, *Mobility*, *Oxygen saturation*, *Nausea*), which are normally the most common observations made by nurses. Many concepts are sparse, with 55 concepts appearing at most five times and 23 concepts appearing only once or twice in the labeled data (e.g., *Drain output*, *Prosthetic use*, *Foreign object removal*).

The provided schema defines 193 unique observation concepts with explicit value types and, for categorical concepts, enumerated allowable values. The schema is dominated by categorical fields (130 SINGLE_SELECT and 12 MULTI_SELECT), with additional NUMERIC (20) and STRING (31) concepts. Categorical concepts have relatively small label sets on average, but include larger option lists for clinically

richer fields such as *Bowel movement description* (12 options) and *Pain severity* (11 options). The schema also includes 15 explicit "unit" concepts (e.g., oxygen saturation unit, respiration unit), representing measurements as value–unit pairs. In the labeled data, 178 out of the 193 schema concepts appear at least once, which supports broad coverage while preserving a realistic distribution.

Overall, this dataset is well-suited for evaluating schema-constrained extraction because it combines (i) conversational nurse-patient transcripts of varying length, (ii) a large and heterogeneous schema with strict type and enumeration requirements, including unit pairing, and (iii) a naturally imbalanced concept distribution that makes rare observations harder to capture while maintaining strong schema-adherence.

4. Methodology

4.1. Overview

Our approach is implemented as a modular workflow in which each component can be enabled or disabled, allowing controlled comparisons across design choices and LLM backbones. Figure 1 provides a complete illustration of our system. Given an input transcript x and the schema $\mathcal{S}$, the system constructs a prompt that conditions the LLM on (i) task instructions enforcing strict structured output, (ii) schema (full or pruned schema), and (iii) retrieved in-context exemplars when RAG is enabled. The LLM is instructed to output only a JSON list of (id, v) pairs (concept identifier and value), which we then deterministically expand to the required submission format by looking up each concept's canonical `name` and `value_type` from the schema; this avoids errors from hallucinated names or types and guarantees metadata adheres to the schema. Concretely, each final predicted observation is represented as `id, name, value_type, value`, where `name` and `value_type` are taken from $\mathcal{S}$ and `value` is generated by the LLM and validated against $\mathcal{S}$. We evaluate the same pipeline configurations with two LLM backbones: *Llama-4-Scout-17B-16E-Instruct* (open-weights) (Meta, 2025) and *GPT-5.2* (closed-weights) (OpenAI, 2025), and report the comparative effects of RAG, schema-constraints (full or pruned schema), and second-pass auditing in the ablation study.

4.2. Retrieval Corpus and Knowledge Base

We use the training split as the sole retrieval corpus and source of in-context demonstrations. Each training instance is parsed from JSONL as a tuple (x_i, y_i) containing a transcript x_i and a gold label observation list y_i. We convert the gold label observations into a standardized representation $y_i^{\text{std}} = [(id_j, v_j)]_{j=1}^{K_i}$, where id_j is the concept identifier and v_j is the corresponding value. We also derive a summary text for each training instance by mapping their gold label concept IDs to their concept names and concatenating them with a delimiter (e.g., "$\text{name}_1 \mid \text{name}_2 \mid \cdots$"). This yields two complementary textual views per training example: the original transcript text and a concept name summary text derived from the schema.

For semantic retrieval, we embed training instances and queries using the same encoder within each pipeline: Sentence-Transformers `all-MiniLM-L6-v2` (Reimers et al., 2021) for the Llama-4 and OpenAI `text-embedding-3-large` (OpenAI, 2024) for the GPT-5.2. We precompute and cache embeddings for each transcript view (and, when retrieval is enabled, for the concept name view as well), along with y_i^{std} for use as few-shot examples. At inference time, the query transcript is embedded under the same encoder and compared against the cached training embeddings to retrieve candidate exemplars.

4.3. Exemplar Retrieval for RAG

When retrieval is enabled, we use an exemplar retrieval strategy that leverages both the narrative content of the transcript and the structured observation concepts associated with each training instance. Each training example is represented by two textual views: (i) the original transcript and (ii) a schema-derived concept-name summary constructed by mapping its gold label concept IDs to their standardized schema names. For a query transcript, we first retrieve a candidate pool using cosine similarity over transcript embeddings, and then re-rank the pool using a weighted score:

$$\text{score(candidate pool)} = w_t.s_t + w_c.s_c + w_l.s_l$$

where s_t is cosine similarity between transcript embeddings, s_c is cosine similarity between concept-name summary embeddings, and s_l is a lexical overlap score computed as Jaccard similarity over token sets (Schütze et al., 2008). We set $(w_t, w_c, w_l) = (0.70, 0.25, 0.05)$ based on experiments on the dev set, as this combination yielded the best overall retrieval quality and downstream extraction performance. The top-k exemplars under this score are inserted into the prompt as few-shot demonstrations, each consisting of the exemplar transcript paired with its standard gold label output as a JSON list of (id, v) pairs. We vary the number of retrieved exemplars with $k \in \{3, 5, 10, 20, 30, 50\}$ and analyze its effect.

4.4. Schema Constraints

We study two strategies for conditioning the LLM on the task schema $\mathcal{S}$, which enumerates all observation concepts and their value constraints. In *full schema* prompting, we provide the complete schema to the model, including each concept's identifier, standard name, value type, and (when categorical) its allowable value set. In *pruned candidate schema* prompting, we instead construct a per-instance candidate concept set $\mathcal{C}(x) \subseteq \mathcal{S}$ intended to cover concepts relevant to transcript x while limiting the output space. The prompt then includes only a compact candidate table for $\mathcal{C}(x)$ (identifier, name, value type, and truncated allowable values for categorical concepts).

To construct $\mathcal{C}(x)$, we score concepts by lexical match between transcript tokens and concept-name tokens (and, for categorical concepts, limited overlap with enumerated tokens), retaining the highest-scoring lexical candidates. We augment this with a semantic schema match by retrieving nearest concept names from an embedding index over all names in the schema and keeping matches above a fixed similarity threshold. We then expand $\mathcal{C}(x)$ using retrieval expansion by adding concept IDs observed in the retrieved exemplars' gold label outputs, and we additionally inject a small set of common observation patterns (e.g., vital signs and frequently occurring nursing assessment fields). Finally, we size the candidate set to a target budget (with minimum and maximum bounds) by trimming low-scoring candidates while preserving a small set of common concepts, or padding with additional lexical/semantic candidates when the set is too small.

4.5. Prompt Construction and LLM Inference

For each input transcript, we construct a prompt designed to elicit strictly structured outputs that adhere to the schema. Figure 2 illustrates the prompt, which begins with system-level instructions that enforce a JSON-only response and explicitly prohibit additional text, code fences, or explanations. We require a single JSON list of objects, each containing exactly an observation identifier and its value (i.e., (id, v)), and we instruct the model to omit any observation not supported by evidence in the transcript and to avoid duplicate identifiers.

We then provide either a full schema or a pruned candidate schema and add few-shot examples when RAG is enabled: up to k retrieved exemplars formatted as a transcript followed by its gold label output. Finally, we append the query transcript and a concise instruction to produce the JSON output. We apply the same prompt template across LLM backbones, differing only in the inference interface

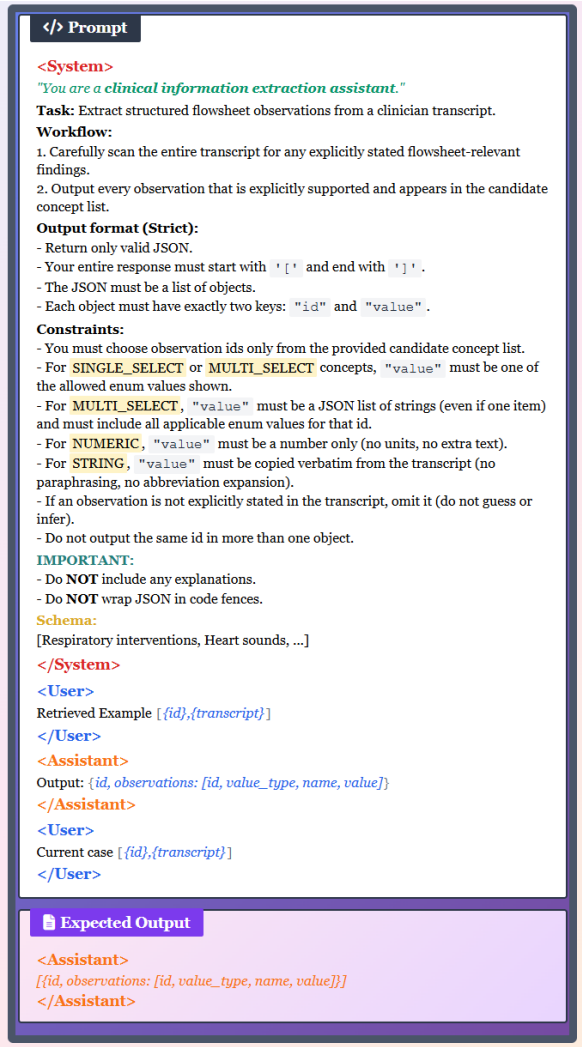


Figure 2: The structured prompt with retrieved exemplars, schema, and the expected output.

(Transformers for *Llama-4-Scout-17B-16E-Instruct* and API-based inference for *GPT-5.2*).

4.6. Postprocessing and Schema Validation

We apply postprocessing to make model outputs robust. First, we deterministically clean and parse the generated text into a JSON list of (id, v) pairs, removing common formatting errors when necessary. We then validate predictions against the schema $\mathcal{S}$ by dropping unknown concept IDs and normalizing values according to each concept's `value_type`. For categorical concepts, outputs are constrained to allowable enumerated values; for multi-select concepts, values are converted to lists and deduplicated; and for numeric concepts, numeric strings are converted to numbers when possible. Duplicate IDs are removed by keeping the first valid occurrence. Finally, we expand each validated (id, v) pair into the required sub-

LLM Backbone	Setting	Precision	Recall	F_1 score
Llama-4	Prompt-only + full schema	72.92	52.84	61.28
Llama-4	Prompt-only + pruned candidate schema	74.64	64.02	68.92
Llama-4	Prompt + RAG + full schema	77.24	68.24	72.46
Llama-4	Prompt + RAG + pruned candidate schema	78.58	68.50	73.20
GPT-5.2	Prompt-only + full schema	81.97	73.17	77.32
GPT-5.2	Prompt-only + pruned candidate schema	69.80	83.57	76.07
GPT-5.2	Prompt + RAG + full schema	78.59	82.03	80.27
GPT-5.2	Prompt + RAG + pruned candidate schema	81.71	73.54	77.41
GPT-5.2	Prompt + RAG + full schema + 2nd pass	78.62	82.18	**80.36**

Table 2: Performance (%) across Llama 4 and GPT-5.2 with schema/RAG configurations. Values are rounded to two decimals. Best F_1 is bolded.

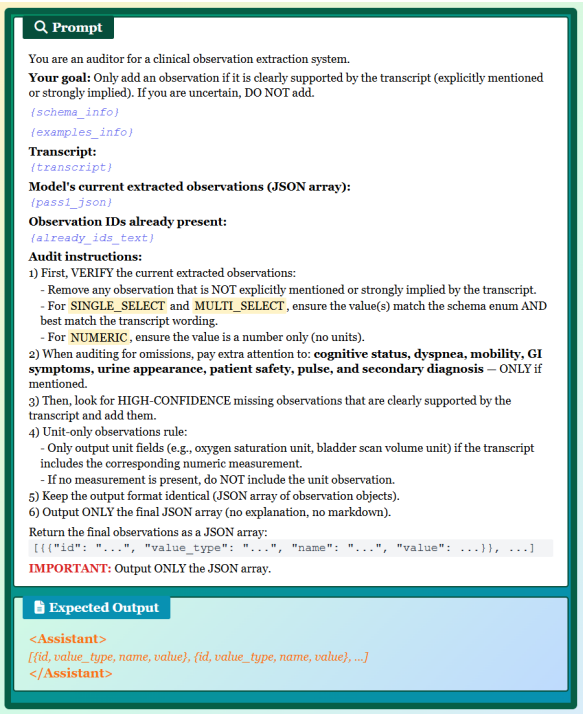


Figure 3: The structured prompt for the second pass audit with retrieved exemplars, schema, first pass solution, and the expected output.

mission format by retrieving the standard `name` and `value_type` from the schema, producing `id`, `name`, `value_type`, `value` entries.

4.7. Second-Pass Auditing

Figure 3 shows the prompt for the second pass auditing step. We evaluate this step as a refinement stage over the first pass, which is the initial system output produced by the LLM from a single pass over the transcript and schema. The auditor is prompted with the original transcript, the same schema used in the first pass (full schema or pruned candidate schema), and the first-pass prediction, and is instructed to remove unsupported items, correct schema adherence issues, and add only clearly supported missing observations. The audited output is then passed through the same schema validation and deterministic expansion step as the first pass.

4.8. Evaluation Strategy

We compute precision, recall, and F_1 over extracted observations by matching predicted and reference items by concept `id` and `value`. For MULTI_SELECT concepts, the evaluation script expands each selected value into a separate item prior to matching, effectively scoring multi-select outputs at the individual choice level. Categorical and string values are compared as strings, and numeric values are compared after conversion to floating point.

5. Results and Discussion

5.1. Overview

Table 2 reports the performance of our retrieval-augmented, schema-constrained extraction pipeline across LLM backbones and design variants. The strongest configuration uses *GPT-5.2* with OpenAI `text-embedding-3-large` for exemplar retrieval, combined with full schema prompting and $top\text{-}k = 30$ retrieved exemplars, yielding the best overall F_1. We observe that increasing the number of retrieved exemplars beyond this point produces only marginal changes, suggesting diminishing returns from additional exemplars. Finally, adding a second-pass auditing stage yields a small additional improvement, indicating that most performance gains are driven by the primary system, with second-pass primarily addressing residual normalization and schema-adherence issues.

5.2. Ablation Study

We conduct an ablation analysis to isolate the contributions of RAG, schema, and second-pass auditing. We start with a prompt-only baseline using *Llama-4-Scout-17B-16E-Instruct* and the full schema, which illustrates the difficulty of extracting and normalizing concepts when the model must reason over a large output space under strict type and enumeration constraints. In this case, errors are dominated by missed concepts and schema-inconsistent outputs, reflecting both limited grounding from the input and the challenge of selecting among many semantically adjacent concepts.

Adding retrieval augmentation improves performance for both LLMs. Concretely, retrieved examples reduce false negatives by providing localized evidence for what constitutes an extractable observation and how it should be normalized under the schema. They also reduce formatting and validity errors by reinforcing the expected JSON structure and value through examples. This effect is especially visible for categorical concepts, where the examples encourage selecting values from the allowed set rather than generating free-form paraphrases.

We then study different schema strategies. For open-weight LLM, pruned candidate schema prompting produces the largest improvement, yielding the best results for *Llama-4* when combined with retrieval. This behavior is consistent with a constrained-decision-space effect: pruning removes many implausible concepts, making it easier for a smaller model to adhere to the schema and to focus extraction on the most relevant portion of the schema. This change improves recall while keeping precision stable or slightly improved, indicating that the model misses fewer relevant observations without a corresponding increase in over-extraction.

For the larger close-weights model, we observe a different interaction between retrieval and schema-constraint. Candidate pruning tends to increase recall but decrease precision, consistent with a checklist-like bias in which the reduced candidate set implicitly encourages broader extraction even when supporting evidence is weak. Under retrieval augmentation, the best *GPT-5.2* performance is instead obtained with full schema prompting, suggesting that a larger model can exploit entire schema context to better disambiguate related concepts while using retrieved exemplars primarily for grounding and normalization. Finally, adding a second-pass auditing stage on top of the best *GPT-5.2* configuration yields a small additional gain. The magnitude of this improvement indicates that most performance is already captured by the combination of retrieval and schema-constraints, with the second-pass primarily correcting residual schema-adherence and normalization errors rather than changing extraction coverage.

Taken together, the ablation results support a clear finding: retrieval augmentation is broadly beneficial, but the optimal schema constraint is model-aware. Pruned candidate schema prompting is critical for controlling the effective output space for the smaller open-weight model, whereas the larger closed model benefits more from full schema context when paired with RAG. The best system performance is because of the combination of (i) exemplar grounding that reduces omissions and stabilizes normalization, and (ii) a full schema that enables consistent disambiguation under strict type and enumeration requirements.

Error type	Count
False positives	**608**
Spurious concept (id not in gold)	498
Wrong value (id in gold)	110
False negatives	**485**
Missed concept (id not predicted)	382
Wrong value (id predicted)	103

Table 3: Breakdown of false positives (FP) and false negatives (FN) into spurious concept errors versus value mismatches.

6. Error Analysis

We analyze errors by comparing predicted observations against the reference output (with MULTI_SELECT values treated as one item per selected choice). Across the test set (199 cases), the system produces 2236 true positives, 608 false positives, and 485 false negatives, as shown in Table 3, indicating that most residual errors arise from either predicting extra items or missing gold label items rather than near-miss value disagreements. Consistent with this, 81.9% of false positives (498/608) correspond to *spurious predictions* whose concept IDs are not present anywhere in the ground truth set for that case, while the remaining 18.1% (110/608) are cases where the correct concept is predicted but with an incorrect value. On the false-negative side, 78.8% (382/485) are *omitted concepts* (the concept ID is not predicted at all), and 21.2% (103/485) are attributable to value mismatches. We also find that errors increase primarily with *clinical content density*: the number of errors correlates more strongly with the number of gold label items ($\rho = 0.45$) and the number of predicted items ($\rho = 0.53$) than with transcript length ($\rho = 0.24$), suggesting that dense dictations amplify both omission risk and over-extraction.

A notable fraction of errors concentrates in a small set of systematic patterns. First, we observe a persistent class of representation-level mis-

Top FP concepts				Top FN concepts			
id	**Name**	**Type**	**FP**	**id**	**Name**	**Type**	**FN**
3	Respiratory interventions	MULTI_SELECT	65	3	Respiratory interventions	MULTI_SELECT	73
31	Secondary diagnosis	STRING	61	148	Gastrointestinal symptoms	MULTI_SELECT	29
0	Broset violence checklist	SINGLE_SELECT	41	0	Broset violence checklist	SINGLE_SELECT	23
162	Patient identification	STRING	26	6	Weightbearing status	MULTI_SELECT	20
6	Weightbearing status	MULTI_SELECT	19	26	Delirium symptoms	STRING	17
26	Delirium symptoms	STRING	17	7	Oral mucosa status	SINGLE_SELECT	16
7	Oral mucosa status	SINGLE_SELECT	17	179	Temperature unit	SINGLE_SELECT	15
117	Patient safety	SINGLE_SELECT	15	89	Mobility	SINGLE_SELECT	15
130	Cognitive status	SINGLE_SELECT	13	107	Fall risk identification	SINGLE_SELECT	14
67	Dyspnea	SINGLE_SELECT	12	130	Cognitive status	SINGLE_SELECT	14

Table 4: Most frequent concept-level errors. False Positive (FP) counts reflect extra predicted items; False Negative (FN) counts reflect missing gold-labeled items (after expanding MULTI_SELECT choices).

matches involving a small subset of low-number concept identifiers that appear with leading zeros in the reference output (e.g., `03` vs. `3`). These mismatches produce both false positives and false negatives even when the extracted content is clinically correct, and they account for a substantial share of total errors (147/485 false negatives and 159/608 false positives). A related phenomenon occurs for temperature units, where the schema's categorical tokens do not match the reference encoding, yielding systematic value mismatches. These issues motivate future work on *robust standardization* for heterogeneous schema and annotations, including normalization of identifier formats and encoding-aware mapping for categorical tokens, as a general requirement when deploying schema-constrained extraction across data sources with inconsistent conventions.

Second, the largest precision losses arise from a small number of *over-predicted* concepts that are plausible given the transcript but rarely used in the references. For example, Table 4 shows that `Secondary diagnosis` appears only 3 times in the gold labels but is predicted 61 times, and `Patient identification` appears only 3 times in the gold labels but is predicted 26 times. This pattern suggests that the model often identifies clinically relevant information but maps it into schema slots that are sparsely annotated or used under stricter, dataset-specific criteria. Addressing this reliably is not a simple rule-based fix; it requires learning concept-specific decision policies and calibration that reflect when a mention should be normalized into a structured field versus left unexpressed.

Finally, residual recall errors frequently reflect *implicit or distributed evidence* that is harder to consolidate into a single normalized observation, especially for checklist-style multi-select fields (e.g., `Gastrointestinal symptoms`) and summary concepts (e.g., `Mobility`). In addition, free-text fields exhibit high brittleness under strict value matching: `Pain description` is present 12 times in gold labels and is frequently predicted, yet none of these predictions match exactly at the string level, indicating that minor paraphrasing or formatting differences can dominate the residual error mass. Improving these cases likely requires methods that better ground string-valued slots in transcript spans (e.g., structured span-copying or alignment) and that model schema overlap explicitly so that evidence expressed in supporting fields is more consistently propagated into the corresponding summary concepts.

7. Conclusion

Schema-constrained extraction from conversational nurse-patient transcripts is an important step toward reducing documentation burden by transforming clinical documentation into structured, schema-adherent observations. Our work presents a modular retrieval-augmented extraction pipeline that combines training-set exemplar retrieval, schema-constrained prompting (full schema or pruned candidate schema), deterministic schema-based postprocessing, and a second-pass audit to refine outputs. Our results support three key findings: pruned candidate schema is beneficial for smaller open-weight models but can be counterproductive for larger models that instead benefit from richer full schema context; retrieval-augmented generation consistently improves performance, with gains strongly influenced by how schema information is presented; and second-pass auditing provides modest additional improvements by correcting residual schema-adherence and normalization issues. Overall, these findings underscore that effective schema-constrained observation extraction is achieved by jointly selecting retrieval and schema-constraint strategies that match the capabilities of the underlying model.

8. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeya Balaji Balasubramanian, Daniel Adams, Ioannis Roxanis, Amy Berrington de Gonzalez, Penny Coulson, Jonas S Almeida, and Montserrat García-Closas. 2025. Leveraging large language models for structured information extraction from pathology reports. *Journal of Pathology Informatics*, page 100521.

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.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Suzhou (China). Association for Computational Linguistics.

Linda Dawson, Maree Johnson, Hanna Suominen, Jim Basilakis, Paula Sanchez, Dominique Estival, Barbara Kelly, and Leif Hanlen. 2014. A usability framework for speech recognition technologies in clinical handover: A pre-implementation study. *Journal of medical systems*, 38(6):56.

Sunyang Fu, David Chen, Huan He, Sijia Liu, Sungrim Moon, Kevin J Peterson, Feichen Shen, Liwei Wang, Yanshan Wang, Andrew Wen, et al. 2020. Clinical concept extraction: a methodology review. *Journal of biomedical informatics*, 109:103526.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Yan Hu, Xu Zuo, Yujia Zhou, Xueqing Peng, Jimin Huang, Vipina K Keloth, Vincent J Zhang, Ruey-Ling Weng, Cathy Shyr, Qingyu Chen, et al. 2026. Information extraction from clinical notes: are we ready to switch to large language models? *Journal of the American Medical Informatics Association*, page ocaf213.

Maree Johnson, Samuel Lapkin, Vanessa Long, Paula Sanchez, Hanna Suominen, Jim Basilakis, and Linda Dawson. 2014a. A systematic review of speech recognition technology in health care. *BMC medical informatics and decision making*, 14(1):94.

Maree Johnson, Paula Sanchez, Hanna Suominen, Jim Basilakis, Linda Dawson, Barbara Kelly, and Leif Hanlen. 2014b. Comparing nursing handover and documentation: forming one set of patient information. *International Nursing Review*, 61(1):73–81.

A H M Rezaul Karim and Ozlem Uzuner. 2025. MasonNLP at MEDIQA-OE 2025: Assessing large language models for structured medical order extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 57–67, Virtual. Association for Computational Linguistics.

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.

Siru Liu, Allison B McCoy, and Adam Wright. 2025. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association*, 32(4):605–615.

Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P Ma, April S Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, et al. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45.

Diwakar Mahajan, Jennifer J Liang, Ching-Huei Tsou, and Özlem Uzuner. 2023. Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. *Journal of biomedical informatics*, 144:104432.

LeAnn Mayer, Dongjuan Xu, Nancy Edwards, and Gordon Bokhart. 2022. A comparison of voice

recognition program and traditional keyboard charting for nurse documentation. *CIN: Computers, Informatics, Nursing*, 40(2):90–94.

AI Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *https://ai. meta. com/blog/llama-4-multimodal-intelligence/, checked on*, 4(7):2025.

George Michalopoulos, Jean-Philippe Corbeil, Cari Bader, Nate Bodenstab, and Asma Ben Abacha. 2026. Overview of the mediqa-synur 2026 shared task on observation extraction from nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

Raghad Muhiyaddin, Asma Elfadl, Ebtehag Mohamed, Zubair Shah, Tanvir Alam, Alaa Abd-Alrazaq, and Mowafa Househ. 2022. Electronic health records and physician burnout: a scoping review. *Informatics and Technology in Clinical Care and Public Health*, pages 481–484.

David Fraile Navarro, Kiran Ijaz, Dana Rezazadegan, Hania Rahimi-Ardabili, Mark Dras, Enrico Coiera, and Shlomo Berkovsky. 2023. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, 177:105122.

OpenAI. 2024. text-embedding-3-large Model. OpenAI API documentation.

OpenAI. 2025. GPT-5.2 Model. OpenAI API documentation.

Nils Reimers, Iryna Gurevych, and Sentence-Transformers contributors. 2021. sentence-transformers/all-MiniLM-L6-v2. Hugging Face model card.

Tiago Rodrigues and Carla Teixeira Lopes. 2025. Harnessing large language models for clinical information extraction: A systematic literature review. *ACM Transactions on Computing for Healthcare*.

Lisa S Rotenstein, A Jay Holmgren, Michael J Healey, Daniel M Horn, David Y Ting, Stuart Lipsitz, Hojjat Salmasian, Richard Gitomer, and David W Bates. 2022. Association between electronic health record time and quality of care metrics in primary care. *JAMA Network Open*, 5(10):e2237086–e2237086.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Darya Shlyk, Tudor Groza, Marco Mesiti, Stefano Montanelli, and Emanuele Cavalleri. 2024. Real: A retrieval-augmented entity linking approach for biomedical concept recognition. In *Proceedings of the 23rd workshop on biomedical natural language processing*, pages 380–389.

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.

Hanna Suominen, Maree Johnson, Liyuan Zhou, Paula Sanchez, Raul Sirel, Jim Basilakis, Leif Hanlen, Dominique Estival, Linda Dawson, and Barbara Kelly. 2015a. Capturing patient information at nursing shift changes: methodological evaluation of speech recognition and information extraction. *Journal of the American Medical Informatics Association*, 22(e1):e48–e66.

Hanna Suominen, Liyuan Zhou, Lorraine Goeuriot, and Liadh Kelly. 2016. Task 1 of the clef ehealth evaluation lab 2016: Handover information extraction.

Hanna Suominen, Liyuan Zhou, Leif Hanlen, and Gabriela Ferraro. 2015b. Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. *JMIR medical informatics*, 3(2):e4321.

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.

Zaifu Zhan, Shuang Zhou, Mingchen Li, and Rui Zhang. 2025. Ramie: retrieval-augmented multi-task information extraction with large language models on dietary supplements. *Journal of the American Medical Informatics Association*, 32(3):545–554.

Night Shift Nerds at MEDIQA-SYNUR 2026: Pushing Small Large Language Model Capability for Clinical Observation Extraction and Normalization from Nurse Dictation using RLVR

Bayu Aryoyudanta[1], Maria Yuliana[1], Mikie Rachman[2], I Made Agus Setiawan[1]
[1]University of Pittsburgh, [2]Washington University in St. Louis
{baa181, mgy9, made.agus}@pitt.edu, rachmanmikie@wustl.edu

Abstract

We presented a small decoder-only language model for clinical observation extraction and normalization from nurse dictation developed using Reinforcement Learning with Verifiable Rewards (RLVR). We fine-tune Qwen3-1.7B model using a two-stage pipeline: (1) supervised fine-tuning (SFT) with an augmented chain-of-thought (CoT) dataset generated by a teacher model to mitigate RL cold-start, followed by (2) GRPO-based RLVR with multi-component reward functions that verify output format, concept presence, value type, and value correctness using the shared-task ontology (193 concepts) as a verifier. On the development set, SFT+GRPO substantially outperforms GRPO-only **(F1 0.803 vs. 0.620)**. After the test holdout was released, our final system achieved **0.700** precision, **0.785** recall, and **0.740** F1. Error analysis shows remaining challenges in concept over-detection and missed concepts, as well as boundary errors in categorical and multi-select value types extraction. Our results demonstrates that small language models can enable accurate, cost-effective, and privacy preserving automated clinical documentation for nurse dictation, supporting scalable deployment in low-resource healthcare settings to reduce nurses' documentation burden.

Keywords: Clinical Extraction Task, LLM, SFT, RLVR

1. Task Overview

Clinical documentation refers to the process of recording patient's conditions and communicating actions, decision and clinical reasoning among members of the care team. Its primary aim is to provide concise, history-rich notes that summarize collected information and support clinical impression, diagnosis, treatment and recommended follow-up (Kuhn et al., 2015). Despite the advantages of clinical documentation, the overwhelming workload associated with creating clinical documentations during patient encounters creates a significant burden among clinicians especially nurses where they spent nearly 30% of their time with EHR documentation which leads to less time and attention to patient interaction (Sasseville et al., 2025).

In response, the rapid advancement of Artificial Intelligence (AI) technologies in modern healthcare has positioned AI medical scribe technology as a promising approach for improving clinical documentation workflow and reducing clinician burden. AI medical scribes listen to clinician-patient interactions and generate not only verbatim transcripts but also structured clinical documentation mapped to existing clinical concepts (Mess et al., 2025). While transcription quality from AI medical scribe system has achieved a strong performance, the extraction and normalization of the clinical observation from these transcripts remain a key limitation.

To address this challenge, the MEDIQA-SYNUR shared task focuses on the extraction and normalization of clinical observation from patient and nurse conversational transcripts and mapping them to a large ontology of clinical concepts. The goal is to reduce nursing burden and improve the accuracy and completeness of patient records by automatically capturing clinically salient information from nurse-patient conversations. MEDIQA-SYNUR shared task is part of the 2026 Workshop on Clinical Natural Language Processing (Clinical NLP).

Given that many healthcare settings, including rural hospitals and organizations in low-resource countries, have limited computational budget and infrastructure, there is a need for AI tools that can work within these constraints. This project examines whether the reinforcement learning technique, especially reinforcement learning with verifiable reward (RLVR) (Lambert et al., 2025) can improve the performance of small large language models (LLM), fewer than 2B parameters, for extracting and normalizing clinical observations from patient-nurse conversation transcripts. By focusing on small models, we aim to make this approach more accessible and applicable to a wider range of clinical settings.[1]

2. Related Work

Extracting clinical observations from biomedical texts with natural language processing (NLP) approach has been a long-standing area of research.

[1]Full disclosure: our team includes organizers from this shared task to our code-repository. All codes for data pre-processing, augmentation, fine-tuning, inference, post-processing, and evaluation are available at: https://github.com/yudanta/nsn-mediqa-synur

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 163–173
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Clinical named entity recognition (Clinical NER), typically formulated as a sequence labeling task has been widely studied for extracting clinical information using rule-based systems, recurrent neural networks and transformer-based models trained for NER downstream task(Neumann et al., 2019; Lee et al., 2020). While sequence labeling approaches are effective at identifying biomedical entities, it does have several limitations particularly with respect to entity type rigidity and value normalization. These models are usually trained on a fixed set of entity types, and normalization to standardized clinical concepts often requires additional post-processing modules which beyond sequence labeling architectures capability.

With the emergence of LLM especially the generative pre-trained model, LLM have been increasingly applied to clinical NER task through prompt engineering, often achieving improved performance over the traditional sequence labelling approach(Hu et al., 2024). In addition, LLM can be fine-tuned for Clinical NER task with instruction type following dataset. For instance, Liu et al 2025 (Liu et al., 2025b), fine-tuned an LLM (Llama 3.1 8B) with supervised fine-tuning (SFT) technique with structured schema (i.e. JSON output format) and achieve 90% exact match accuracy witch less annotated data compared to the traditional machine learning approach. Although, SFT-based LLM demonstrated improved schema aware generation and better contextual understanding compared to sequence modelling, SFT primarily teach models to generate outputs given input and output from its training tokens. As a result, SFT-trained model often struggle with unseen cases or scenarios requiring logical verification, which are common in clinical concept extraction task. For instance, the clinical entity might explicitly mentioned but negated or it may need to be inferred implicitly from the surrounding context. These limitations constrain the generalizability of SFT-based LLM for clinical entity extraction and normalization task.

Reinforcement learning (RL) has recently emerged as a powerful paradigm for improving the capability of LLM. For tasks with with verifiable outcomes, (Guo et al., 2025), introduced Group Relative Policy Optimization (GRPO) an RL method that eliminates the need for a critic network by estimating advantages from grouped response comparisons from verifiable reward functions. In addition, their work also demonstrates that RL can substantially improve model reasoning capabilities. Building on this foundation, Tulu 3 (Lambert et al., 2025) introduced Reinforcement Learning with Verifiable Rewards (RLVR), a multi-stage post-training framework that combines SFT with subsequent RLVR training pipeline, achieving state of the art results and empirically outperforming SFT only approaches. In the biomedical domain, Fleming-R1(Liu et al., 2025a) applies RLVR with GRPO for medical reasoning tasks with three complementary innovations: 1) reasoning-oriented data curation, 2) Chain-of-Thought (CoT) cold start training before RL training and 3) two-stage RLVR training focus on different reasoning skills sets data resulting in near-parity performance with GPT-4o model. Building upon these prior works, our projects extend the RLVR framework for clinical observation extraction and normalization task from patient-nurse conversational transcripts.

Field	Content
ID	14
Transcript	*[Clinician]* Alright, let's see here...uh, the patient's high risk for falls. They have a, uh, history of falls and, uh, joint deformity, so, uh, we're providing, um, moderate assistance with mobility...
Observations	Cognitive status: Alert with general confusion; ...; Patient safety: Yes

Table 1: Example of clinical documentation entry from our dataset.

3. Dataset

The 2026 MEDIQA-SYNUR shared-task dataset https://sites.google.com/view/mediqa2026/mediqa-synur aimed to extract medical observation from nurse dictation data. This dataset consists of patient-nurse dictation and normalized extracted clinical information mapped into predefined clinical schema. The patient-nurse dictations were generated through a controlled multi-agent simulation pipeline, and the gold-standard annotations were produced by expert nurses using an open-source, large scale clinical schema (193 concepts) (Corbeil et al., 2025). The dataset is partitioned into training, development and test subsets. In this manuscript, we mainly focus on train and development subsets of MEDIQA-SYNUR dataset. Examples of nurse-patient transcript and its extracted clinical information and clinical concept map can bee seen from Table 1 and 2 while dataset statistics and its clinical observation distribution can be found in Table 3.

4. Methods

We extend the RLVR framework to the task of extracting and normalizing clinical observations from patient - nurse conversational transcripts and mapping them to a large ontology of clinical concepts.

Field	Value
ID	3
Name	Respiratory interventions
Value Type	MULTI_SELECT
Value Enum	• raising head of bed • turn cough • deep breathe • incentive spirometer usage

Table 2: Example clinical concept map for respiratory interventions.

Statistic	Train	Dev	Test
Total examples	122	101	199
Avg. obs. per example	13.8	13.0	12.8
Min. obs. per example	6	7	4
Max. obs. per example	34	33	39
Single-select obs.	1200	951	1853
Multi-select obs.	195	157	344
String obs.	108	70	257
Numeric obs.	182	137	98

Table 3: Summary statistics of the MEDIQA-SYNUR dataset across training, development, and test splits.

Our training pipeline employs a cold start CoT fine-tuning stage followed by RL with GRPO as illustrated in Figure 1. To construct the CoT dataset, we leverage a larger open-source model to generate reasoning traces that aid model's exploration capability by learning concept identification and value extraction through CoT dataset before GRPO training.

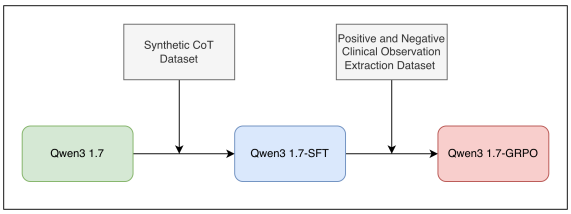


Figure 1: Current Approach Training Pipeline.

4.1. Dataset

The raw dataset from the 2026 MEDIQA-SYNUR shared task was processed to align with the requirements for SFT and RLVR model fine tuning. Each A clinical concept will be paired with a clinical transcript as a context with an output of normalized extracted value. These pair of dataset will be sampled and used as a seed for generating CoT dataset for the SFT fine-tuning process.

4.1.1. Dataset pre-processing and augmentation

The train datasets from this shared-task underwent several pre-processing steps, focusing on dataset formating and negative example augmentation. We created two classes of data: positive and negative examples. The positive dataset consist of consist of nurse dictation transcript paired with clinical concept and its expected extracted value. The Negative dataset consist of nurse dictation transcript paired with clinical concepts that do not appear in the transcript with a null extracted value.

During negative dataset generation, we also generate negative pairs of similar or potentially confusing clinical concepts to model distinguish between related concepts that are not present in the transcript. To support this process, we first created a mapping of clinical concepts and their related or commonly confused concepts as shown in Table 4. Examples of positive and negative samples are shown in Table 5 and Table 6. Summary statistics of the training dataset are presented in Table 7

Concept	Similar / related concepts
Blood pressure	Mean arterial pressure (MAP)
Mean arterial pressure (MAP)	Blood pressure

Table 4: Example of similar clinical concept for blood pressure.

4.1.2. Synthetic dataset augmentation for unseen concepts

During our exploratory data analysis, we found out that the train dataset examples doesn't cover all clinical concepts from the provided ontology. Those unseen clinical concepts from train dataset are: **'Bladder scan volume unit', 'Emesis volume unit', 'Hearing aid', 'Assistance with toileting', 'Seizure activity', 'Excoriated tissue', 'Voice quality', 'Prosthetic use', 'Food allergies', 'Foreign object removal', 'Pupil size', 'Speech content', 'Stool consistency', 'Caloric intake unit', 'Temperature source', 'Tracheostomy status', 'Vaginal discharge', 'Ear examination', 'Cranial nerve function', 'Voiding function', 'Body art', 'Tube feeding type', 'Patient belongings', 'Broset violence checklist - physically threatening', 'Chest expansion', 'Positioning frequency', and 'Motor reflexes'**. To overcome this unseen concept, we generated synthetic datasets to fill the unseen concept gaps in our dataset. During the

Field	Content
Transcript	[Clinician] Patient presents with moderate dyspnea. Uh, they're using accessory muscles—um, you can see the increased effort with their breathing. We've been, uh, managing this with a few respiratory interventions...
Concept	Dyspnea
Value Type	SINGLE_SELECT
Value Enum	['None', 'Mild', 'Moderate', 'Severe']
Value	Moderate

Table 5: Example of positive train dataset.

Field	Content
Transcript	[Clinician] Patient presents with moderate dyspnea. Uh, they're using accessory muscles—um, you can see the increased effort with their breathing. We've been, uh, managing this with a few respiratory interventions...
Concept	Emesis color
Value Type	SINGLE_SELECT
Value Enum	['dark green', 'green']
Value	None

Table 6: Example of negative train dataset.

process, we utilize OpenAI GPT-4o to generate this synthetic examples.

4.1.3. CoT data augmentation

In addition to positive and negative examples for model fine-tuning, we also generated a CoT dataset to support the SFT training process. The CoT dataset examples were generated using OpenAI GPT OSS 120B as teacher models. As seed for generation, we sampled the pre-processed dataset decribed in Section 4.1.1. The output of this process is a collection of CoT dataset consisting of nurse dictation transcript, clinical concepts and their associated reasoning process as illustrated in Table 8. At the end of CoT data augmentation process, we selected 3066 examples for SFT training as summarized in Table 9

4.2. SFT with CoT

In this study, to address the cold start problem and improve overall model performance as described in RLVR training pipeline(Lambert et al., 2025), we first fine-tuned the Qwen3 1.7B base model with the CoT dataset generated in Section 4.1.3. This

Category	Count
Total clinical concepts covered	193
Negative samples	5818
Positive samples	3036
SINGLE_SELECT	6109
STRING	1147
NUMERIC	939
MULTI_SELECT	659

Table 7: Summary of the training dataset used for model development.

CoT-initialized model is expected to enhance base model exploration capability during the reinforcement learning process, thereby mitigating the cold start issue and improving overall task performance.

4.3. RLVR Fine-Tuning

During GRPO training, instead of using answer correctness as a single verification mechanism, we employ multi-component reward functions that verify several aspects of the model output:

- **Response format**: verifies that the response contain both reasoning and answer section enclosed within <think></think> and <answer></answer> tags.
- **Presence of clinical concept**: verifies whether the target clinical concept is present in the patient-nurse transcript or not.
- **Correctness of clinical concept's value type**: verifies that the extracted value type matches the concept value type defined in the clinical concept ontology.
- **Correctness of extracted clinical value**: verifies that the extracted value matches the gold standard annotation values.

In this experiment we employed non-binary reward values, which typically assigning **+1** for correct response and **(0 or -1)** for incorrect responses, to provide more granular reward control and penalties based on model behavior. The combination of these reward functions and non-binary reward values is intended to address the unique challenges of this clinical observation, extraction, normalization and ontology mapping where errors can occur in multiple aspects.

In our experiment, we iteratively adjusted reward and penalty values for clinical concept detection based on error analysis to balance false positive and false negative clinical concept detection. Because our developed model tended to produce more false positive observations, we set higher

Field	Content
Transcript ID	160
Concept	Heart rate
Value type	NUMERIC
Label	Positive
Ground-truth value	88
Transcript	[Clinician] Patient is, uh, currently in a pos...
CoT reasoning	The transcript contains the exact phrase **“Heart rate is being monitored at 88 beats per minute”** (line… “Heart rate is being monitored at 88 beats per minute, with heart sounds…”). This wording directly mentions the clinical concept “Heart rate,” satisfying the requirement that the concept be explicitly named in the text. Because the phrase supplies a numeric measurement (“88”) together with the unit “beats per minute,” it unambiguously provides the quantitative value for the heart-rate concept, and no other heart-rate numbers appear elsewhere in the note. Consequently, the correct extracted value for the “Heart rate” concept is **88**, matching the ground-truth specification. Given the clear, verbatim statement and the absence of any contradictory data, we are highly confident (near-certain) that the heart-rate value is 88 bpm.

Table 8: Example instance from the augmented chain-of-thought (CoT) dataset used for supervised fine-tuning.

Category	Count
Total clinical concepts covered	193
Negative samples	2042
Positive samples	1024
SINGLE_SELECT	2091
STRING	379
NUMERIC	338
MULTI_SELECT	258

Table 9: Summary of the augmented chain-of-thought (CoT) dataset used for supervised fine-tuning.

penalty to false positive cases compared to false negative cases. Using non-binary reward values provided greater flexibility in steering model behavior during RL training.

In this study, we performed a full-parameter model fine-tuning for both SFT and RLVR training pipeline by considering that our experiment is limited to small LLM with the base models constrained to maximum of 2 billion of model parameters. Both SFT and RLVR training process were implemented using Huggingface's TRL library.

4.4. Model Evaluation and Performance Metrics

Following the evaluation from the 2026 MEDIQA-SYNUR shared task in clinical extraction, we will employ **Precision**, **Recall** and **F1** evaluation metrics to assesses the performance of the proposed approach. The evaluation script can be access through this link: https://github.com/abachaa/MEDIQA-SYNUR-2026/blob/main/mediqa_synur_eval_script.py

5. Results

Table 10 presents the evaluation metrics across training configurations. Our model gained substantial advantages from the SFT training with the augmented CoT dataset compared to the model trained solely using GRPO-RL, which align with training recipe suggested in prior work from(Lambert et al., 2025; Guo et al., 2025). Furthermore, the accumulated reward for the model trained on SFT + GRPO-RL was higher than trained directly using GRPO-RL. As shown in Figure 2 the model initialized with SFT starts with higher reward values, helping overcome the cold-start problem commonly observed in RLVR model training.

By the time the test holdout was published, our fine-tuned model achieved a **precision of 0.700**, **recall of 0.785** and **F1-scores of 0.740**.

6. Error Analysis

To better understand the strengths and limitations of our developed system, we conducted an error analysis on the **development (dev) dataset** outputs from our best-performing model. Overall, the mean number of clinical observations per transcript in the development dataset is 13.019, while the model predicts an average of 14.316 observations,

Method	Dataset	Precision	Recall	F1-score
GRPO-only	Train	0.496	0.828	0.621
	Dev	0.494	0.830	0.620
SFT+GRPO (Best Model)	Train	0.772	0.832	0.801
	Dev	0.770	0.838	0.803

Table 10: Models (Qwen3-1.7B) performance comparison.

Figure 2: Overall reward function between GRPO-RL only and SFT-Followed by GRPO-RL Training.

indicating that our model tends to extract slightly more concepts than present in the reference annotations. On average, 2.772 extracted observations per transcript correspond to false positives. While, our model correctly extracts a mean of 10.762 observations. Among these correctly identified observations, the mean number of value extraction errors is 0.782, suggesting that our fine-tuned model is generally effective at extracting and normalizing clinical observation values although some errors remain.

6.0.1. Medical Concept Detection

Further error analysis on medical concept detection shows that our fine-tuned model still struggles with both over-detection (false positives) and missed observation (false negatives). Across development dataset, our model produced a total of 280 false-positive clinical observations, indicating an over-extraction tendency, where clinical concepts are predicted even when they are not present in the transcript. These false-positive errors frequently occurred in symptoms or assessment related concepts such as: gastrointestinal symptoms, fall risk identification, cognitive status as well as unit-based concepts such as oxygen saturation unit, heart rate unit, and bladder scan volume unit. This patterns suggest that our model sometimes infer plausible observations from transcript but the real observation is not exist. Examples for both false-positive and false-negative observation can be seen from Table 11.

In the other hand, our fine-tuned model also produced a total of 66 false-negative observations, where clinical concepts actually present in the transcript but not detected by our model. The most commonly missed clinical concepts including secondary analysis, patient safety, mobility, orientation and patient identification. These false-negative cases indicate that our model still lacks robustness in detecting clinical concepts, even when they are expressed explicitly in the transcript. Taken together, this medical concept detection analysis suggests that while our model demonstrates strong overall clinical concept detection, it still shows a tendency toward both over and under detection, which warrant further improvements.

6.0.2. Extracted Concept Values

Error analysis on extracted values shows that among correctly extracted concepts, 97 instances contain incorrect values, with most errors occurring in SINGLE_SELECT (40) and STRING (19), followed by MULTI_SELECT (16) and a smaller number of NUMERIC errors (4). For MULTI_SELECT clinical concepts, our model often over or under select options, for instance predicting "dark" and "cloudy" observation where only "cloudy" is annotated. For SINGLE_SELECT concept, our model show grading or label-boundary confusion such in predicting "moderate" instead of "severe" in Dyspnea concept. For STRING type value, model often extract more phrase compared to the annotated value for instance our model extract "2+ bilaterally" instead of "2+" for Peripheral pulse concept. For NUMERIC type value, error occurs in numeric values that is stated in a compound form, for example is blood pressure observation which written as 160/95mmHg. Overall these pattern suggest that model still need improvement primarily for MULTI_SELECT observation and capability in categorical boundary decision.

7. Conclusion

We presented a fine-tuned decoder only LLM for clinical observation extraction based on Qwen3-1.7B trained on the MEDIQA-SYNUR shared-task dataset using an SFT followed by GRPO-RL pipeline. In addition to the shared-task dataset,

we incorporated an augmented CoT dataset during SFT, which improved model performance after GRPO-RL training.

Our experimental results suggest that clinical observation detection, extraction, and normalization given a predefined clinical concepts ontology can be formulated as a verifiable problem, making RLVR suitable for this task. The clinical concept ontology can serve as a verifier for the RLVR reward functions, enabling the model to learn reasoning behaviors that improve overall task performance, even for relatively small LLM such as Qwen3-1.7B. Therefore, this resource-efficient model is not only accessible, but also cost-effective and privacy-preserving, making it suitable for low-resource setting and scalable to broader clinical settings to help reduce nurse burden in clinical documentation.

8. Future Work

Error analysis highlights that our developed model still suffer from false positive and false negative observations, as well as incorrect value extraction for correctly identified clinical concept from patient-nurse dictation data. These limitation warrant future study to improve clinical concept detection, value extraction and normalization from patient-nurse dictation data.

In addition to improving RLVR reward function design with ontology-aware and concept-level penalties to better balance both false-positive and false-negative detection behaviors, new approaches such as interactive agentic AI system that support clarification for ambiguous clinical concepts are wort exploring.

9. Ethical Consideration

Failures in LLM-based clinical extraction system could exacerbate existing health disparities and increase the risk of patient safety. Given that the LLM may exhibit demographic biases and lower accuracy in non-English language, a systematic extraction errors could affect certain populations. In addition, missing critical observations like confusion or pain assessment for patients could create or further increase clinical documentation disparities that may affects to individuals healthcare outcomes.

For user data privacy, the 2026 MEDIQA-SYNUR shared task datasets were generated through multi-agent interaction in closed environment without using any real person dataset. Therefore, issues that might arise related to personal health information (PHI) can be minimized.

10. Acknowledgment

We gratefully acknowledge Health and Rehabilitation Informatics (HARI) Lab hari.pitt.edu for providing computational resources and infrastructure support for the model fine-tuning experiments.

11. Bibliographical References

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Yu-Cheng Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025. Empowering Healthcare Practitioners with Language Models: Structuring Speech Transcripts in Two Real-World Clinical Applications. ArXiv:2507.05517.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan,

Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

Thomson Kuhn, Peter Basch, Michael Barr, Thomas Yackel, and for the Medical Informatics Committee of the American College of Physicians*. 2015. Clinical Documentation in the 21st Century: Executive Summary of a Policy Position Paper From the American College of Physicians. *Annals of Internal Medicine*, 162(4):301–303.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. ArXiv:2411.15124.

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.

Chi Liu, Derek Li, Yan Shu, Robin Chen, Derek Duan, Teng Fang, and Bryan Dai. 2025a. Fleming-R1: Toward Expert-Level Medical Reasoning via Reinforcement Learning. ArXiv:2509.15279 version: 1.

Longchao Liu, Long Lian, Yiyan Hao, Aidan Pace, Elaine Kim, Nour Homsi, Yash Pershad, Liheng Lai, Thomas Gracie, Ashwin Kishtagari, Peter R. Carroll, Alexander G. Bick, Anobel Y. Odisho, Maggie Chung, and Adam Yala. 2025b. Human level information extraction from clinical reports with finetuned language models. *Scientific Reports*, 15(1):45239.

Sarah A. Mess, Alison J. Mackey, and David E. Yarowsky. 2025. Artificial Intelligence Scribe and Large Language Model Technology in Healthcare Documentation: Advantages, Limitations, and Recommendations. *Plastic and Reconstructive Surgery - Global Open*, 13(1):e6450.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Maxime Sasseville, Farzaneh Yousefi, Steven Ouellet, Florian Naye, Théo Stefan, Valérie Carnovale, Frédéric Bergeron, Linda Ling, Bobby Gheorghiu, Simon Hagens, Samuel Gareau-Lajoie, and Annie LeBlanc. 2025. The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review. *Healthcare*, 13(12):1447.

12. Appendices

Field	Content
False Positive Example	
ID	87
Transcript	[Clinician] Alright, let me gather my notes here for the patient assessment. So, um, we have a patient who's showing, uh, let's see, a mix of urinary and gastrointestinal symptoms. The patient reported, uh, vomiting, and, uh, the emesis was, uh, recorded at about 500 cc. It's that dark green color, which could suggest, um, some gastrointestinal distress or maybe something going on there, like a blockage. [Clinician] On the urinary side, um, the patient is experiencing difficulty urinating and, uh, has this urgency. So, yeah, it might be, um, a sign of some bladder irritation or maybe obstruction, possible stone there. Uh, mobility-wise, the patient is slightly limited, um, but continent, which is good. Uh, this limited mobility could be due to discomfort or pain from these issues. [Clinician] Now, um, checking the oral mucosa, it's dry, which could hint at dehydration, especially with the vomiting and possibly not taking enough fluids. Uh, vital signs are stable-ish, I'd say, with an oxygen saturation at 92% and a heart rate of 85 bpm. So, yeah, this paints a picture of acute distress from these gastrointestinal and urinary issues, not critical but definitely needs attention and maybe, um, further diagnostics to pinpoint the exact cause.
Gold Concept	None (Gastrointestinal symptoms not annotated)
Predicted Concept	Gastrointestinal symptoms
Explanation	The model incorrectly inferred the higher-level concept "Gastrointestinal symptoms" from the mention of vomiting. However, vomiting is annotated under its own concept in the ontology, leading to a false-positive concept detection.
False Negative Example	
ID	10
Transcript	[Clinician] Alright, so, uh, here's what we've got. This patient, elderly, has been, um, experiencing a few issues over the last 24 hours. They're currently, uh, bedridden due to a recent surgery, which limits their, uh, mobility quite a bit. Now, uh, during the assessment, some things stood out. There's evidence of a, uh, constipation episode, which we can see with the perineal edema present. It's, um, making the patient quite uncomfortable. [Clinician] Now, the patient is also dealing with, uh, persistent nausea. They are retching but, uh, haven't had any actual vomiting, so it's more of a gastrointestinal distress without the vomiting. Skin assessment shows, uh, tented skin turgor, indicating possible dehydration. This might, uh, be linked to the constipation they're experiencing.[Clinician] Cardiovascular-wise, the heart rate's a bit high, sitting at around 100 bpm. This could be, um, partially due to dehydration and maybe some pain. The patient, uh, reports pain at about 6 out of 10 and is also having, um, painful swallowing, which complicates, uh, oral intake. Clinician] Overall, it looks like a multifactorial approach is needed here. We'll need to, uh, address hydration management, constipation relief, and, uh, control the pain effectively.
Gold Concept	Mobility
Predicted Concept	Not detected
Explanation	Although the transcript explicitly describes reduced mobility ("bedridden" and "limits mobility"), the model failed to detect the Mobility concept, resulting in a false-negative observation.

Table 11: Examples of model errors from the development dataset. The first example illustrates a false-positive concept detection caused by ontology-level confusion, while the second shows a false-negative detection despite explicit mention in the transcript.

Prompt Template for SFT Training

```
Extract medical concept: {medical_concept_name}
Value Type: {value_type}
Allowed values (choose ONE): {value_enums}

Transcript: {patient_nurse_dictation_transcript}

Select exactly one value from the allowed list.

Expected output format:
<think>
reasoning process
</think>

<answer>
{"concept_present": true/false,
"concept": "concept name",
"value": "value" or null if concept_present is false}
</answer>

{completions}
```

Figure 3: Prompt template for SFT Training.

GRPO Fine-Tuning Example for Clinical Concept Extraction

System Prompt

```
You are a medical concept extraction assistant analyzing conversations between patients and clinician or nurses. Your task is to identify and extract key medical concepts including symptoms, conditions, medications, treatments, procedures, and relevant clinical observations.

1. First, analyze the conversation in the <think> section, identifying:

  • Medical terminology and clinical concepts
  • Patient-reported symptoms and concerns
  • Clinician or nurse assessments and recommendations
  • Temporal relationships and context
  • Whether the target concept is present, absent, or ambiguous
  • For the target concept: locate the specific value or measurement
  • For unit concepts (e.g., 'Heart rate unit'): verify the base measurement is present before extracting the unit

2. Then, provide structured extracted concepts in the <answer> section.

Output format:
<think>
reasoning process
</think>

<answer>
{"concept_present": true/false,
"concept": "concept name",
"value": "value" or null if concept_present is false}
</answer>

IMPORTANT: Set concept_present to true ONLY if a specific value for the concept can be identified in the transcript. Never extract a unit concept (e.g., 'Temperature unit', 'Heart rate unit') unless the corresponding measurement itself is documented.
```

User Prompt

```
Extract medical concept: {medical_concept_name}
Value Type: {value_type}
Allowed values (choose ONE): {value_enums}
Transcript: {patient_nurse_dictation_transcript}
Select exactly one value from the allowed list.
```

Figure 4: Example instruction pair used during GRPO fine-tuning for clinical concept extraction from nurse dictation transcripts.

Extracting Medication Instructions from Dutch General Practice Electronic Health Records with Local Natural Language Processing

Marya Dukmak, Constanza L Andaur Navarro, Artuur M Leeuwenberg
University Medical Center Utrecht, Utrecht University,
Heidelberglaan 100, 3508 GA Utrecht, The Netherlands
maryadukmak@gmail.com, candaurn@umcutrecht.nl, aleeuw15@umcutrecht.nl

Abstract

The extraction of structured medication prescription data from unstructured clinical text remains a critical challenge for clinical research and data standardization. This study investigates the application of Natural Language Processing (NLP) techniques to Dutch electronic health records (EHRs) from the Julius General Practitioners Network. The goal is to automatically extract key prescription attributes including dosage, duration, and medication unit and prepare them for integration into the ConcePTION Common Data Model, to support scalable pharmacoepidemiological research. We compare a lightweight rule-based system with transformer-based models (RobBERT and MedRoBERTa) under the technical constraints of a Trusted Research Environment, where external resources and cloud-based solutions are restricted. Using a dataset of 1,819 manually annotated records, the approaches are evaluated on predictive performance and computational costs. Results show that the rule-based system achieves strong accuracy and computational costs for structured patterns, while transformer-based models demonstrate greater robustness to linguistic variability. However, both approaches encounter difficulties with ambiguous dosage formats and long treatment durations. Our findings indicate that NLP methods can substantially improve the structuring of Dutch prescription data and support scalable pharmacoepidemiological research. Future work should focus on improving generalization and expanding annotated datasets to enhance model reliability.

Keywords: Dutch Electronic Health Records (EHRs), NLP, Medication information extraction

1. Introduction

Electronic Health Records (EHRs) have become a fundamental component of modern healthcare, enabling the large-scale collection of patient data and supporting clinical decision-making as well as medical research. However, a substantial portion of EHR data remains unstructured in free text form, limiting its usability for secondary use. An important secondary use of EHR data is the conduct of pharmacoepidemiological studies that investigate medication safety after their introduction into healthcare practice. Critically, these large-scale (often international) studies rely on common data models (CDMs): data from all participating healthcare centers - with often different local recording practices - are first mapped to the same data structure and semantics, to ensure comparability and usability for research. This article investigates to what degree different natural language processing (NLP) methods can extract relevant medication information from EHR free text, to make them available for pharmacoepidemiological research. Specifically, we focus on the extraction of medication prescription instructions from Dutch general practice EHRs.

1.1. Case study setting

Having accurate data about what medication is prescribed to each patient is essential for pharmacoepidemiological studies that evaluate treatment strategies and medication safety. One of the most used common data models to encode such information used in this field is the ConcePTION CDM (Thurin et al., 2022). We focus on three essential ConcePTION CDM variables in this study: dosage, duration, and medication units. Figure 1 illustrates an example of the input and outputs for the NLP task discussed in the current study.

A second important aspect of this setting is the use of (non-English) Dutch GP EHR data from the Julius General Practitioners Network (Smeets et al., 2018), which has important implications for what NLP techniques can be used: they should handle Dutch clinical free text (with very specific terminology and abbreviations). And, importantly, the use of these (pseudonimized) sensitive data is only permitted in a strict privacy protected digital research environment (DRE): data may not be send to third parties (e.g., proprietary LLM providers) and computational resources are relatively limited. In other words, NLP techniques are constrained to locally runnable arguably smaller models that can handle Dutch clinical text: these are thus the focus of this study.

1.2. Research contributions

Within this case study, we compare the following NLP methods: a rule-based approach with regular expressions, and two Dutch transformer-based models: RobBERT and MedRoBERTa.nl. Our research questions pertain to how well each of these methods extracts the relevant prescription informa-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 174–182
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tion and what is the computational costs.

This paper makes the following contributions:

- We present an NLP pipeline for extracting key prescription attributes dosage, duration, and medication unit from Dutch primary care EHR data.
- We provide a comparative evaluation of rule-based and transformer-based approaches within the constraints of a secure digital research environment.
- We analyze the strengths and limitations of both approaches in terms of performance, efficiency, and robustness to prescribing variability.

By improving the structuring of prescription data, this work supports the development of scalable and reusable clinical datasets, ultimately facilitating large-scale pharmacoepidemiological research.

2. Related Work

Medical information extraction aims to transform large volumes of unstructured clinical text into structured data suitable for analysis. As the adoption of Electronic Health Records has increased, automated extraction methods have become essential for efficiently identifying clinically relevant information (Ford et al., 2016).

Early approaches relied primarily on rule-based systems that leverage handcrafted grammars, lexicons, and domain knowledge. These methods have demonstrated strong performance in structured clinical settings. For example, (Mykowiecka et al., 2009) developed a rule-based system for extracting attributes from Polish clinical documents, achieving precision and recall above 99%. Despite their reliability and interpretability, rule-based approaches often struggle with scalability and require continuous manual updates to capture linguistic variation.

To address these limitations, machine learning techniques such as Conditional Random Fields (CRFs) and Support Vector Machines have been introduced for clinical information extraction. (Patrick and Li, 2009) presented a hybrid system combining rule-based preprocessing with statistical classifiers, achieving an F-score of 93.5% for medication extraction from emergency department notes. Similarly, systems integrating probabilistic models with linguistic rules have shown that combining domain expertise with data-driven methods can improve extraction performance.

More recently, transformer-based architectures have substantially advanced clinical NLP. Pretrained language models enable the capture of contextual relationships within complex clinical narratives while reducing the need for manual feature engineering (Yang et al., 2020). These models have been successfully applied to organize electronic medical records and support clinical decision-making (Landolsi et al., 2022). However, challenges remain, particularly regarding domain adaptation and the handling of specialized medical terminology (Tayefi et al., 2021).

While medication information extraction methods with transformer models have recently been explored for non-English (Fabacher et al., 2025; Campillos-Llanos et al., 2025), the majority of studies and transformers primarily utilized English corpora, and these results or models cannot be directly generalized to Dutch medical text without evaluation or adaptation. Research on Dutch clinical NLP remains comparatively limited(AlShuweihi et al., 2021). Addressing this gap, (Muizelaar et al., 2024) evaluated BERT-based models for extracting lifestyle characteristics from Dutch clinical notes, demonstrating the effectiveness of domain-specific pre-training.

A key resource in this area is MedRoBERTa.nl, the first transformer-based language model trained specifically on Dutch electronic health records (Verkijk and Vossen, 2021). The model outperforms general-purpose Dutch language models on several clinical NLP tasks, highlighting the importance of in-domain language resources for underrepresented languages.

Together, these studies indicate a shift from deterministic extraction toward context-aware language models, while also emphasizing the continued relevance of interpretable approaches in clinical environments. Building on this body of work, the present study compares rule-based and transformer-based strategies for extracting structured medication information from Dutch primary care EHRs.

3. Methodology

3.1. Data and annotation

The data used in this study originates from the Julius General Practitioners Network (JGPN) (Smeets et al., 2018), a large primary care database made available for medical research, covering over 390 general practitioners in the Utrecht region, spanning from years 1919 to 2024, and containing over three million medication records.

Data were first filtered for recency (after year 2000) and availability of the values for the relevant columns (prescription instruction and product strength). Then, from the remaining 861,515 records, a random subset of 2,000 prescription records was selected to represent real-world clinical

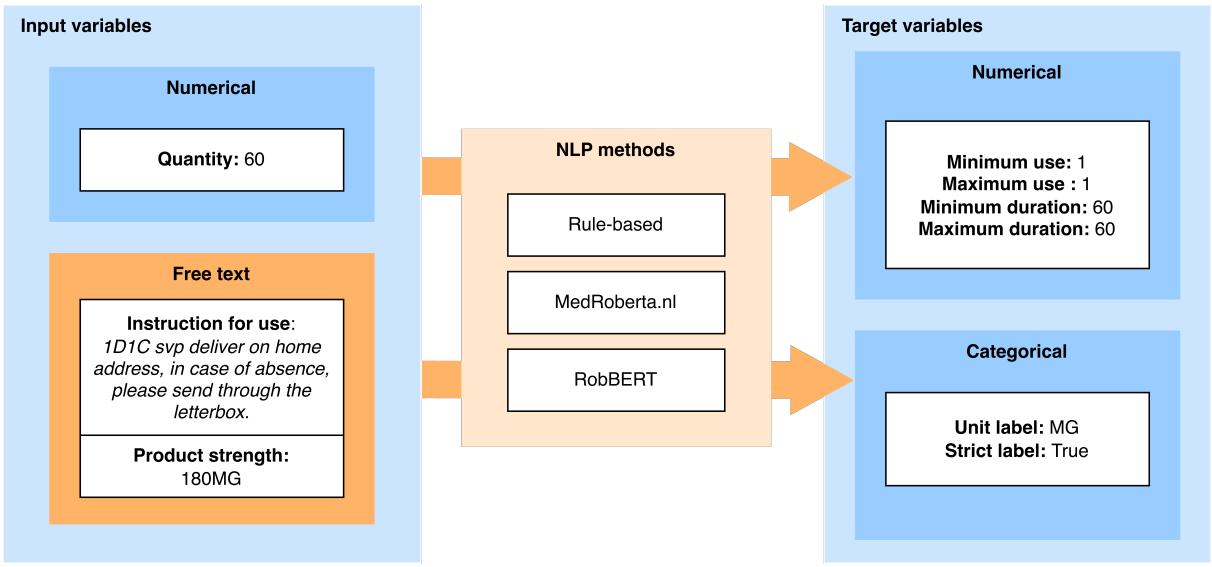


Figure 1: Schematic input output overview for an (artificial, English-language) example.

documentation practices. The subset includes both frequently occurring and rare dosage patterns, and no data balancing was applied in order to preserve the natural distribution of prescription instructions.

Using LabelStudio (Tkachenko et al., 2020–2025) each record was first annotated by a single annotator (Dutch native, non-clinical, data scientist) for key medication attributes: dosage, duration, and medication unit (descriptions found in Table 1. Any uncertain cases were held out and checked with a pharmacist or general practitioner, depending on the source of uncertainty. A random subset of 100 records was annotated by a second annotator (Dutch native, non-clinical, data scientist) to calculate inter-annotator agreement, which can be found in Table 2. A screenshot of the annotation tool and the LabelStudio configuration are added to the supplementary materials.

Finally, after excluding cases the pharmacist and GP could not label (e.g., sometimes interpretation required the full patient journey instead of only the prescription instruction), 1,819 were selected to develop and evaluate the NLP models.

3.2. Experimental setup

The 1,819 annotated records were split in a train (80%) and test set (20%) at the GP practice level (to prevent leakage from having the same GP, with likely more similar documentation style in the train and test set). For each of the three modeling methods, the train set was used for its development (also for the rule-based system) and test set for its evaluation.

3.2.1. Evaluation

All models were evaluated using the test set.

Model performance was assessed using standard information extraction metrics, including accuracy, precision, recall, and macro-averaged F1-score. For numerical targets, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were reported to capture prediction deviations.

Computational cost was measured as CPU runtime (seconds), CPU memory peak (in Mb), and GPU memory peak (in Mb).

This study evaluates two extraction strategies for structuring medication prescription instructions: a rule-based approach and transformer-based models. Both methods were designed to identify three target attributes dosage, duration, and medication unit from free-text prescription records.

3.2.2. Models and computational setup

All models used the variables: *prescription* (free text), *product strength* (free text) and *quantity* (numerical) as their input to extract the six target variables. While *quantity* is not free text, it is needed to arrive at the duration: duration can be calculated by dividing the prescribed quantity by the daily dosage (often present in the prescription field).

Rule-based model. The rule-based system was implemented using handcrafted patterns and regular expressions to capture structured dosage expressions. The rules were iteratively refined to accommodate common prescribing formats, including numerical ranges and abbreviated notation.

Transformer-based models. To complement the rule-based model, transformer-based language models were evaluated for their ability to generalize across linguistic variation. We employed two Dutch-language models: RobBERT, a general-domain RoBERTa-based model, and MedRoBERTa.nl, a

Variable name	Description	Type
Min quantity	Minimum quantity of the prescribed units per day	Numerical
Max quantity	Maximum quantity of the prescribed units per day	Numerical
Min duration	Minimum duration (in days) of the prescribed treatment	Numerical
Max duration	Maximum duration (in days) of the prescribed treatment	Numerical
Strict label	Whether the prescription was strict (e.g., not strict could be "as needed")	Categorical
Unit label	The unit used to indicate the strength of the medication (e.g., milligrams, milliliters)	Categorical

Table 1: Overview of the six target variables

Column	n compared	Agreement rate (%)	Cohen's Kappa κ	Mean absolute difference
Min quantity	100	87.85	0.767	0.150
Max quantity	100	86.92	0.767	0.159
Min duration	100	85.85	n/a	4.632
Max duration	100	85.85	n/a	4.934
Strict label	100	98.13	0.490	n/a
Unit label	100	87.62	0.517	n/a

Table 2: Inter-annotator agreement for 100 examples. Mean absolute differences are reported in units/day for quantity fields and in days for duration fields.

domain-specific model pre-trained on Dutch electronic health records.

Using the training set, each model was fine-tuned in both a multitask (MT) setting (sharing weights of the transformer, to jointly output all target labels at once) and a per-target (PT) setting (separate transformer weights per target label).

Models were fine-tuned using AdamW (learning rate of $1x10^{-}5$, and weight decay of 0.01), batch size 32, using mixed precision (PyTorch AMP), with a single Nvidia T4 GPU) for 30 epochs. Cross-entropy loss and Huber loss ($\delta = 10$) were used for categorical and numerical outcomes respectively. For categorical outcomes, to address the high class imbalance, class weights were applied during training. For the numerical targets, as the distributions were very left-skewed, a log transformation was used (inverted at evaluation).

Computational setup Experiments were conducted with the following allocated resources: AMD EPYC 7V12 with 4 vCPUs (1thread/core), Nvidia Tesla T4 with 16GB VRAM (2.7GB used during measurement), Driver 535.247.01, CUDA 12.2, a x86_64 (64-bit), Little Endian, BogoMIPS 4890, Full VM (Hyper-V).

4. Results

This section presents the performance of the rule-based and transformer-based approaches on the extraction of medication prescription target attributes. The comparison focuses on extraction performance and computational costs.

Overall, the rule-based system achieved the highest extraction accuracy on structured prescription patterns, confirming its effectiveness in scenarios where expressions closely match predefined rules. However, its performance declined when encountering unseen abbreviations or irregular dosage formats that were not covered by the handcrafted patterns. Transformer-based models demonstrated greater robustness when processing complex or composite instructions and were better able to handle off-pattern phrasing.

Both approaches showed reduced performance on specific prescription types. In particular, long treatment durations were frequently underestimated, and fine-grained unit distinctions (e.g., similar lexical forms) led to classification errors. Complex and system-generated dosage patterns were also prone to misinterpretation.

From a computational perspective, the rule-based method was substantially more efficient, requiring minimal runtime (0.02 seconds) and memory (112.7 MB) while operating without GPU resources. Transformer models, in contrast, demanded significantly greater computational capacity, particularly during training.

5. Discussion

This study provides a comparative evaluation of rule-based and transformer-based approaches for extracting structured medication information from Dutch primary care EHRs. The results demonstrate that both strategies offer distinct advantages, and their suitability largely depends on the intended deployment context.

The rule-based system delivered consistently strong performance, particularly for duration prediction and unit classification. Its deterministic design enables stable numerical extraction when prescription patterns follow recognizable structures. In addition, the minimal computational requirements make

Model	Strict label				Unit label			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Rule-based	0.98	0.95	0.99	0.97	0.9	0.76	0.87	0.81
RobBERT (multitask)	0.99	0.95	0.99	0.97	0.82	0.74	0.78	0.74
MedRoBERTa (multitask)	0.99	0.97	0.99	0.99	0.79	0.76	0.73	0.71
RobBERT (per target)	0.99	0.97	0.99	0.98	0.83	0.78	0.74	0.73
MedRoBERTa (per target)	1	1	1	1	0.86	0.8	0.77	0.76

Table 3: Extraction performance on the test set for categorical targets.

Model	Min quantity		Max quantity		Min duration		Max duration	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Rule-based	0.12	0.03	1.2	0.18	0.43	0.04	8.16	1.07
RobBERT (multitask)	0.29	0.36	0.38	0.33	2377.8	33.5	1386.1	25.17
MedRoBERTa (multitask)	0.2	0.28	0.29	0.27	1479.7	25	1409.4	24.23
RobBERT (per target)	0.06	0.17	0.16	0.2	4382.9	43.7	1673.4	27.3
MedRoBERTa (per target)	0.02	0.1	0.17	0.17	1692	26.3	1508.9	24.6

Table 4: Extraction performance on the test set for numerical targets.

Model	**Split**	**Runtime [s]**	**CPU peak [MB]**	**GPU peak [MB]**
Rule-based	Test	0.02	112.7	–
RobBERT (MT)	Train	312.76	1423.9	2271.2
	Test	0.10	1524.0	1134.0
MedRoBERTa (MT)	Train	176.38	1506.0	2443.4
	Test	0.11	1521.0	1203.0
RobBERT (PT)	Train	961.46	2522.9	6743.0
	Test	0.10	1894.0	1034.0
MedRoBERTa (PT)	Train	986.65	2683.8	7255.5
	Test	0.14	1521.0	1203.0

Table 5: Average runtime (CPU time in seconds) and peak memory use (MB) over 5 runs.

it especially attractive for secure research environments where external dependencies and hardware resources may be limited.

Transformer-based models, in contrast, showed greater adaptability to linguistic variability and complex prescription phrasing. However, this flexibility came at the cost of increased computational demand and reduced robustness in extracting certain numerical targets.

The observed performance differences emphasize an important trade-off between interpretability and adaptability. Rule-based systems provide transparent decision logic that facilitates error tracing and supports clinical trust, whereas transformer models operate largely as black boxes despite their predictive capabilities. In healthcare settings, where explainability is often a prerequisite for adoption, this distinction remains highly relevant.

Another key finding concerns model stability. While transformer models achieved near-perfect classification scores, they exhibited higher variance in duration predictions, particularly for extreme treatment lengths. This suggests that numerical extrapolation remains challenging for data-driven approaches trained on limited samples.

Taken together, these findings indicate that rule-based and transformer-based methods should not necessarily be viewed as competing alternatives. Instead, they may function as complementary components within hybrid extraction pipelines.

5.1. Limitations

Several limitations should be considered when interpreting these results. First, the annotated dataset is relatively small which may impact generalization, and was not split temporally, which could lead optimism in the evaluation estimates with regard to recent samples. Second, the data originate from a single regional general practitioner's network, potentially limiting portability to regions with different prescribing conventions. Finally, transformer models were evaluated without extensive data augmentation or advanced optimization strategies, leaving room for model improvements.

5.2. Implications for Clinical NLP

From a practical perspective, these results highlight the importance of aligning model selection with operational constraints. Lightweight systems may provide immediate value in resource-constrained environments, whereas transformer-based approaches become increasingly attractive as larger annotated datasets become available.

More broadly, this study contributes empirical evidence from the underrepresented domain of Dutch clinical NLP, supporting ongoing efforts to expand language resources beyond English-centric research.

6. Conclusion

This paper presented a comparative study of rule-based and transformer-based methods for extracting structured medication prescription information from Dutch primary care EHRs. The results demonstrate that rule-based extraction remains a highly reliable and computationally efficient solution for structured prescribing patterns, while transformer models provide greater flexibility in handling linguistic variation.

Rather than identifying a single superior approach, our findings highlight the complementary strengths of deterministic and data-driven methods. Selecting an appropriate strategy therefore depends not only on predictive performance but also on operational constraints such as computational resources, interpretability requirements, and deployment context.

By providing empirical evidence from the relatively underexplored domain of Dutch clinical NLP, this work contributes to the development of scalable approaches for structuring prescription data and supports the broader goal of improving data readiness for large-scale clinical research.

Future work should focus on expanding annotated datasets, improving robustness for numerical prediction tasks, and exploring hybrid architectures that combine the stability of rule-based systems with the adaptability of transformer models.

Acknowledgments

We thank Wijnand Laan for his consultations on the JGPN data processing, Nicole Boekema-Bakker for setting up and managing the digital research environment, and Sander van Doorn and Boris de Graaff for their help in the annotations.

Ethical statement

All data in this study were processed in accordance with the GDPR-compliant procedures of the JGPN, and the protocol was reviewed in consultation with a privacy officer of Utrecht University.

Data availability

Due to the privacy sensitivity of the data they could not be shared. Code used to conduct the experiments is publicly available on GitHub: https://github.com/MariaDukmak/UMC-medication-extraction-submission-/tree/main.

References

Mohamed AlShuweihi, Said A. Salloum, and Khaled Shaalan. 2021. *Biomedical Corpora and Natural Language Processing on Clinical Text in Languages Other Than English: A Systematic Review*, pages 491–509. Springer International Publishing.

Leonardo Campillos-Llanos, Ana Valverde-Mateos, and Adrián Capllonch-Carrión. 2025. Hybrid natural language processing tool for semantic annotation of medical texts in spanish. *BMC bioinformatics*, 26(1):7.

Thibaut Fabacher, Erik-André Sauleau, Emmanuelle Arcay, Bineta Faye, Maxime Alter, Archia Chahard, Nathan Miraillet, Adrien Coulet, and Aurélie Névéol. 2025. Efficient extraction of medication information from clinical notes: an evaluation in 2 languages. *Journal of the American Medical Informatics Association*, 32(12):1855–1864.

E. Ford, J. Carroll, Helen E. Smith, D. Scott, and J. Cassell. 2016. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association : JAMIA*, 23:1007 – 1015.

Mohamed Yassine Landolsi, Lobna Hlaoua, and Lotfi Ben Romdhane. 2022. Information extraction from electronic medical documents: State of the art and future research directions. *Knowledge and Information Systems*, 65(2):463–516.

Hielke Muizelaar, Marcel Haas, Koert Van Dortmont, Peter Van Der Putten, and Marco Spruit. 2024. Extracting patient lifestyle characteristics from Dutch clinical text with bert models. *BMC Medical Informatics and Decision Making*, 24(1).

Agnieszka Mykowiecka, Małgorzata Marciniak, and Anna Kupść. 2009. Rule-based information extraction from patients' clinical data. *Journal of Biomedical Informatics*, 42(5):923–936.

Jon Patrick and Min Li. 2009. A cascade approach to extracting medication events. *ResearchGate*.

Hugo M Smeets, Marlous F Kortekaas, Frans H Rutten, Michiel L Bots, Willem van der Kraan, Gerard Daggelders, Hanneke Smits-Pelser, Charles W Helsper, Arno W Hoes, and Niek J de Wit. 2018. Routine primary care data for scientific research, quality of care programs and educational purposes: the julius general practitioners' network (jgpn). *BMC health services research*, 18(1):735.

Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtliebsen. 2021. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews Computational Statistics*, 13(6).

{Nicolas H} Thurin, Romin Pajouheshnia, Giuseppe Roberto, Caitlin Dodd, Giulia Hyeraci, Claudia Bartolini, and et al. 2022. From inception to conception: Genesis of a network to support better monitoring and communication of medication safety during pregnancy and breastfeeding. *Clinical Pharmacology and Therapeutics*, 111(1):321–331.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020–2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Stella Verkijk and Piek Vossen. 2021. MedRoBERTa.nl: A language model for Dutch electronic health records. *Computational Linguistics in the Netherlands Journal*, 11:141–159.

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942.

Supplementary Materials

Listing 1: Label Studio template (part 1)

```
<View>
  <Header value="[Annotatie] JHN Medicatie
      Tabel (100 voorbeelden)" style="font-size:
      20px; margin: 10px 0;"/>

  <!-- Informatie uit dataset -->
  <View style="border: 2px solid #ccc; border-
      radius: 10px; padding: 15px; margin: 15px 0;
       background-color: #f9f9f9;">
    <Header value="[Informatie uit dataset]" style
      ="font-size: 16px; margin-bottom: 10px;"/>
    <Text name="id_text" value="[ID] $Id"/>
    <Text name="voorschrijfdatum_text" value="[
      Voorschrijfdatum] $Voorschrijfdatum"/>
    <Text name="einddatum_text" value="[
      Einddatum] $Einddatum"/>
    <Text name="hoeveelheid_text" value="[
      Hoeveelheid] $Hoeveelheid"/>
    <Text name="product_sterkte_text" value="[
      Product Sterkte] $Product_sterkte"/>
    <Text name="gebruik_text" value="[
      Gebruiksvoorschrift] $Gebruiksvoorschrift"/>
  </View>

  <!-- Gebruikvoorschrift Details -->
  <View style="border: 2px dashed #9ad7df;
      border-radius: 10px; padding: 16px; margin:
       10px 0 20px; background-color: #eaf7fb;">
    <Header value="[Gebruikvoorschrift Details]"
      style="font-size: 16px; margin-bottom: 8px;
      "/>
    <View style="font-size: 13px; line-height: 1.4;
       margin-bottom: 10px;">
      <Text name="instructions" value="- Min/Max
       presc per day: Beschrijf in getallen (bv. '1-2
       tablet per dag'  min 1, max 2)."/>
      <Text name="instructions2" value="- Min/
      Max presc duration days: Geef cijfers op
      basis van hoeveelheid en gebruiksvoorschrift
      ."/>
      <Text name="instructions3" value="- Strict /
      Not Strict: Strict betekent dat dit nauw moet
      worden opgevolgd; anders waar 'zo nodig'
      staat of / niet strict."/>
    </View>
  </View>
</View>
```

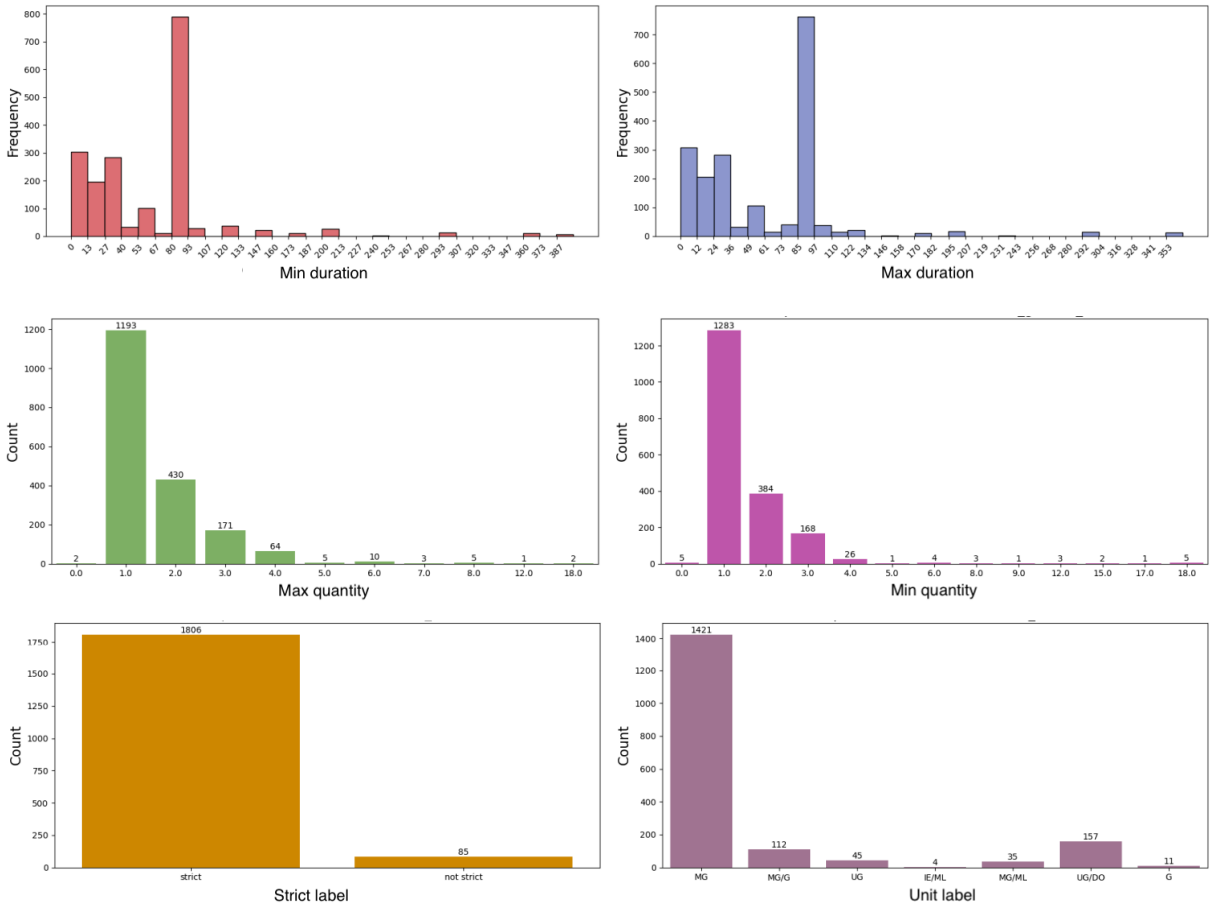


Figure 2: Visualisation of the target label distributions.

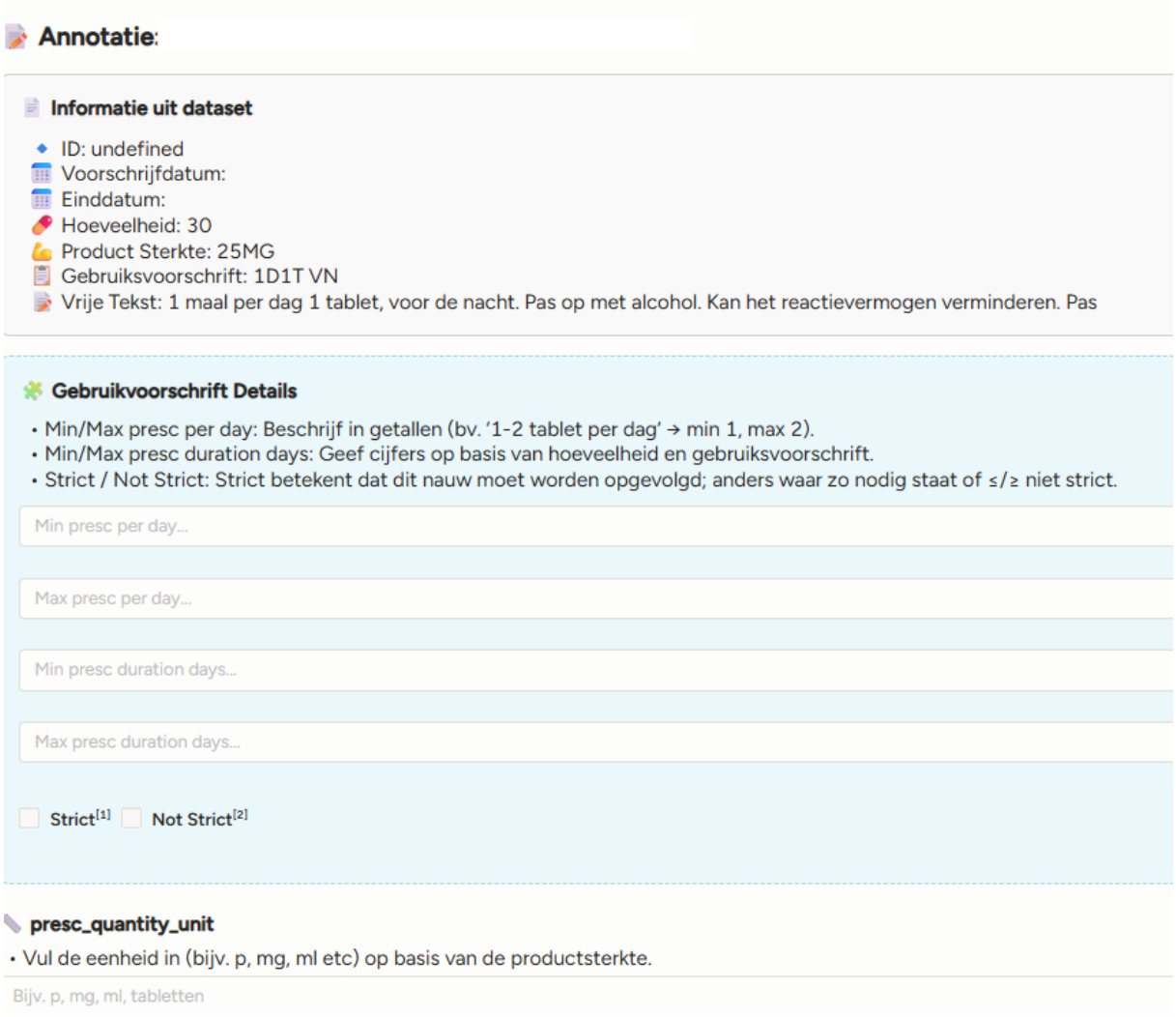


Figure 3: A screenshot of the annotation environment (for an artificial example).

Contextual Clinical Extraction: Integrating Foundation Models with Domain-Specific Validation Rules

Ankit Singh, Siva Satyanarayana Raju
Team Gladiators MEDIQA-SYNUR @ ClinicalNLP / LREC 2026
singhankit16@gmail.com, pusapati.badri@gmail.com

Abstract

We present a hybrid system for schema-guided clinical information extraction from synthetic nursing dictations, developed for the MEDIQA-SYNUR shared task at ClinicalNLP 2026. Our approach combines a large language model (Claude Opus 4) with comprehensive prompt engineering, multi-layered post-processing pipelines, and rule-based extraction using over 400 domain-specific regex patterns. The system extracts structured observations across 193 clinical concepts spanning vital signs, neurological assessments, respiratory status, and functional evaluations. On the development set, our system achieves 74.27% precision, 84.32% recall, and 78.98% F1 score. Detailed error analysis reveals that secondary diagnosis over-extraction (32 false positives) and cognitive status classification (12 value errors) represent primary challenges. We provide comprehensive methodology documentation to enable reproducibility, including prompt templates, validation rules, and regex patterns. We discuss limitations including cost-performance tradeoffs of proprietary LLMs and deployment considerations for healthcare settings.

Keywords: clinical NLP, information extraction, large language models, nursing documentation, schema-guided extraction, MEDIQA, hybrid systems

1. Introduction

Clinical documentation represents a significant burden in modern healthcare, with nurses spending substantial time on administrative tasks that detract from direct patient care. Studies estimate that nurses spend up to 35% of their shift on documentation (Baumann et al., 2018), and the transition to Electronic Health Records (EHRs) has not reduced this burden as initially hoped (Joukes et al., 2018). Voice-based documentation systems offer potential efficiency gains by allowing clinicians to dictate observations naturally, but converting these narratives into structured EHR entries requires sophisticated natural language understanding.

The MEDIQA-SYNUR shared task (Michalopoulos et al., 2026) addresses this challenge by focusing on extracting structured clinical observations from synthetic nursing dictation transcripts. The task presents several technical challenges: (1) a large schema space of 193 clinical concepts across diverse domains; (2) four distinct value types (NUMERIC, STRING, SINGLE_SELECT, MULTI_SELECT) requiring different extraction strategies; (3) strict ontological constraints where extracted values must exactly match enumerated options; and (4) the need to distinguish explicitly stated observations from implicit or speculative information.

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in clinical NLP tasks, including medical entity recognition (Agrawal et al., 2022), clinical question answering (Singhal et al., 2023), and medical text summarization. However, direct application of LLMs to structured clinical data extraction faces challenges including hallucination of plausible but incorrect observations, inconsistent output formatting, and failure to adhere to strict ontological constraints (Ji et al., 2023). These issues are particularly concerning in clinical settings where extraction errors could affect patient care.

We propose a hybrid approach that combines the reasoning and generalization capabilities of foundation models with the precision and reliability of rule-based validation systems. Our contributions include:

(1) A comprehensive prompt engineering strategy incorporating complete schema specification, few-shot learning with diverse training examples, and domain-specific extraction guidelines for challenging clinical concepts.

(2) A multi-layered post-processing pipeline with six specialized filters addressing common LLM extraction errors, including speculative diagnosis removal, physiological range validation, and context-based filtering.

(3) A rule-based extraction component employing over 400 contextually-aware regex patterns organized by clinical domain, with validation mechanisms to prevent false positives.

(4) An intelligent merging strategy that balances LLM comprehensiveness with rule-based precision, along with detailed error analysis quantifying system performance across error types and clinical concepts.

2. Related Work

2.1 Clinical Information Extraction

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 183–191
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Clinical information extraction has a rich history spanning rule-based, statistical, and neural approaches. Early systems relied heavily on medical ontologies and hand-crafted rules. MetaMap (Aronson and Lang, 2010) maps biomedical text to UMLS concepts using lexical and syntactic analysis, achieving high precision for concept identification but requiring extensive manual curation. The clinical Text Analysis and Knowledge Extraction System (cTAKES; Savova et al., 2010) provides a comprehensive pipeline for clinical NLP including named entity recognition, assertion classification, and relation extraction, built on rule-based and machine learning components. These systems established important baselines but struggled with the linguistic variability of clinical narratives.

Negation detection represents a critical subtask in clinical NLP, as clinicians frequently document absent findings. The NegEx algorithm (Chapman et al., 2001) uses simple regular expressions to identify negation triggers and their scope, achieving strong performance on discharge summaries. Extensions like ConText (Harkema et al., 2009) broadened scope to include temporality and experiencer identification. Our system incorporates similar negation handling through both rule-based pattern matching and LLM prompt instructions.

2.2 Transformer-Based Clinical NLP

The transformer architecture (Vaswani et al., 2017) revolutionized clinical NLP through domain-specific pretraining. BioBERT (Lee et al., 2020) pretrained BERT on PubMed abstracts and PMC full-text articles, demonstrating improved performance on biomedical NER and relation extraction. ClinicalBERT (Alsentzer et al., 2019) further adapted BERT to clinical text using MIMIC-III clinical notes, showing gains on clinical NLP benchmarks. PubMedBERT (Gu et al., 2021) demonstrated that pretraining exclusively on domain-specific text outperformed mixed-domain approaches, achieving state-of-the-art results on multiple biomedical NLP tasks.

However, these encoder-based models require task-specific fine-tuning and struggle with large schema spaces like the 193 concepts in MEDIQA-SYNUR. Fine-tuning separate classifiers for each concept would be impractical and would not leverage cross-concept dependencies. Our approach addresses this by using a generative LLM that can handle the full schema in a single prompt.

2.3 Large Language Models for Clinical Tasks

The emergence of large language models has opened new possibilities for clinical NLP. Brown et al. (2020) demonstrated that GPT-3 could perform few-shot learning across diverse NLP tasks without fine-tuning. Subsequent work applied this capability to clinical domains: Agrawal et al. (2022) showed that LLMs are effective few-shot clinical information extractors, achieving competitive performance with minimal task-specific examples. Singhal et al. (2023) demonstrated that LLMs encode substantial clinical knowledge and can perform medical question answering at expert level.

However, hallucination remains a significant concern for LLM-based extraction (Ji et al., 2023). LLMs may generate plausible but incorrect clinical observations, particularly when prompted to extract information that is not explicitly stated. This is especially problematic for secondary diagnoses where the model may infer conditions from symptoms. Our post-processing pipeline specifically addresses this through speculative language filtering and schema validation.

2.4 Schema-Guided Extraction

Schema-guided extraction constrains model outputs to conform to predefined ontologies. Rastogi et al. (2020) introduced the Schema-Guided Dialogue dataset, demonstrating that providing schema descriptions improves zero-shot generalization to new domains. Chen et al. (2021) extended this to multi-domain dialogue state tracking. Our approach adapts schema-guided principles to clinical documentation, where ontological constraints are stricter and extraction errors have higher stakes.

Hybrid approaches combining neural models with rule-based post-processing have shown promise in clinical NLP. Wang et al. (2018) surveyed clinical information extraction applications, noting that the best-performing systems often combine multiple approaches. Our system follows this paradigm, using LLM extraction for broad coverage and rule-based validation for precision.

3. Task Description and Dataset

3.1 MEDIQA-SYNUR Shared Task

The MEDIQA-SYNUR shared task (Michalopoulos et al., 2026) focuses on extracting structured clinical observations from synthetic nursing dictation transcripts. The task simulates a real-world scenario where nurses verbally document patient assessments, and an NLP system must convert these dictations into structured EHR entries conforming to a clinical ontology.

The dataset was generated through a controlled multi-agent simulation pipeline to ensure diverse and clinically realistic scenarios. Gold-standard annotations were produced by expert nurses using an open-source, large-scale clinical ontology. The dataset comprises 122 training samples, 101 development samples, and a held-out test set for final evaluation. Each sample contains a transcript field with natural

nursing dictation and an observations field containing structured annotations.

3.2 Clinical Observation Schema

The observation schema defines 193 clinical concepts organized across multiple clinical domains. Table 1 summarizes the schema distribution by domain and value type. The domains span the full scope of nursing assessment, from vital signs and neurological status to functional capacity and safety considerations.

Domain	Concepts	Example Fields
Vital Signs	15	O2 sat, HR, BP, temp, MAP
Neurological	22	Orientation, GCS, pupils, cognition
Respiratory	18	Dyspnea, breath sounds, O2 delivery
Cardiovascular	14	Rhythm, edema, JVD, pulses
Gastrointestinal	16	Bowel sounds, nausea, vomiting
Genitourinary	12	Urine output, appearance, odor
Musculoskeletal	18	Motor strength, mobility, ROM
Skin/Wounds	20	Turgor, wounds, pressure injuries
Safety	15	Fall risk, bed alarm, restraints
Functional	25	ADL assistance, transfers, gait
Other	18	Pain, nutrition, IV therapy
Total	193	-

Table 1: Distribution of clinical concepts by domain in the SYNUR schema.

Each observation concept is associated with one of four value types, each requiring different extraction and validation strategies:

NUMERIC: Quantifiable measurements such as vital signs (heart rate, oxygen saturation, temperature) and volumes (urine output, emesis). These require extraction of numeric values and validation against physiologically plausible ranges.

STRING: Free-text observations such as pain descriptions, wound characteristics, and patient identification. These require minimal normalization but may contain verbose LLM outputs that need cleaning.

SINGLE_SELECT: Categorical observations with mutually exclusive enumerated options, such as dyspnea severity (None/Mild/Moderate/Severe) or cognitive status. These require exact matching against the schema's allowed values.

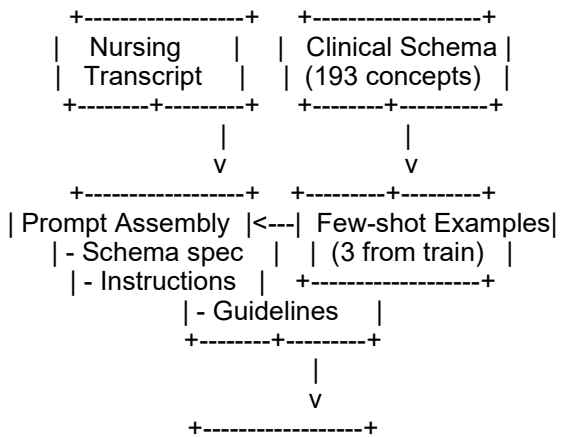
MULTI_SELECT: Observations allowing multiple concurrent values from an enumerated set, such as breathing patterns (labored, shallow, irregular) or respiratory interventions. These require parsing of conjunctions and list structures.

3.3 Evaluation Metrics

The official evaluation uses precision (correct predictions / total predictions), recall (correct predictions / total gold annotations), and F1 score. A prediction is considered correct only if both the observation ID and value exactly match the gold standard. For MULTI_SELECT fields, the extracted values must match as sets (order-independent). Numeric values are compared with type coercion (integer vs. float).

4. System Architecture

Our hybrid extraction system consists of three main components operating in sequence: (1) an LLM-based primary extractor using Claude Opus 4 with comprehensive prompt engineering; (2) a multi-layered post-processing pipeline with domain-specific filters and corrections; and (3) a rule-based supplementary extractor with contextual validation. Results from both extraction methods are merged through an intelligent strategy and validated against the clinical schema. Figure 1 illustrates the complete pipeline.



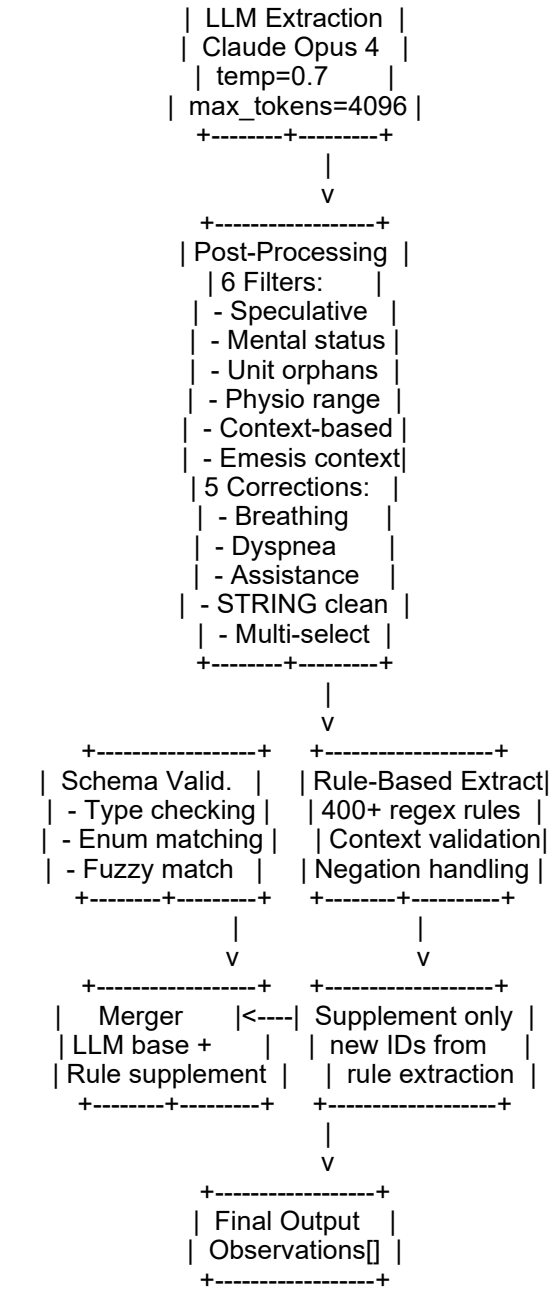


Figure 1: System architecture showing the complete hybrid extraction pipeline with all processing stages.

4.1 LLM-Based Extraction

We use Claude Opus 4 (model ID: claude-opus-4-5-20251101) as the primary extraction engine. The model is accessed via the Anthropic API with the following configuration: temperature=0.7 for balanced consistency and flexibility, and max_tokens=4096 constraining output length only. The input prompt, including schema specification, few-shot examples, and transcript, fits within the model's context window without truncation.

The prompt engineering strategy incorporates five key components designed to maximize extraction accuracy while minimizing hallucination:

Schema Specification: The complete schema of 193 observation concepts is provided in the prompt, including concept IDs, names, value types, and enumerated options for SELECT fields. This explicit schema grounding helps the model understand the extraction targets and valid output values.

Few-Shot Examples: We include 3 training samples as few-shot examples, selected to demonstrate diverse observation types. Examples are chosen to cover: (a) multiple value types in a single transcript, (b) challenging concepts like multi-select respiratory interventions, and (c) proper handling of negated or absent findings.

Extraction Rules: The prompt explicitly instructs the model to: (a) extract only observations that are explicitly stated in the transcript; (b) avoid speculation or inference from symptoms; (c) preserve exact wording for STRING values; (d) map to schema-compliant values for SELECT types; and (e) handle negated findings appropriately.

Domain-Specific Guidelines: Special instructions address challenging clinical concepts: (a) respiratory observations require explicit dyspnea severity mentions rather than inference; (b) cognitive status should distinguish between alert, confused, and disoriented states; (c) secondary diagnoses should only include confirmed conditions, not differentials.

Multi-Select Handling: Instructions for parsing conjunctions ('nausea and vomiting') and list structures ('raising head of bed, turn cough, deep breathe') to extract multiple values for MULTI_SELECT fields.

4.2 Post-Processing Pipeline

The LLM output undergoes extensive post-processing through six specialized filters and five correction mechanisms. These address systematic extraction errors identified through iterative development on the training and development sets.

4.2.1 Filtering Mechanisms

Speculative Diagnosis Filter: Removes secondary diagnosis extractions containing uncertain language markers: 'possibly', 'could be', 'might', 'suspected', 'likely', 'probably', 'suggestive of', 'consistent with', 'differential', 'rule out', 'r/o', 'versus', 'signs of'. This prevents extraction of differential diagnoses as confirmed conditions.

Mental Status Validation: The 'forgets limitations' mental status value is only retained if explicitly mentioned in the transcript. General forgetfulness or confusion does not trigger this extraction, preventing over-generalization from related cognitive findings.

Unit Field Orphan Removal: Removes unit fields (e.g., 'heart rate unit') when the corresponding value field was not extracted. This addresses LLM tendency to extract partial observation pairs.

Physiological Range Validation: Numeric values are validated against clinically plausible ranges: heart rate 40-200 bpm, respirations 8-40/min, oxygen saturation 50-100%, temperature 32-43 C, mean arterial pressure 40-150 mmHg. Values outside these ranges are filtered.

Context-Based Filtering: Cardiac values (heart rate) require cardiac context keywords within 50 characters ('heart rate', 'pulse', 'HR', 'bpm'). Respiratory values require respiratory context ('respiration', 'RR', 'breaths per minute'). This prevents numeric values from unrelated contexts being incorrectly mapped.

Emesis Context Filter: Prevents extraction of urine appearance values (e.g., 'dark') when the transcript context discusses emesis rather than urinary output. Triggered by emesis keywords ('vomit', 'emesis') in proximity to the extracted value.

4.2.2 Correction Mechanisms

Breathing Pattern Correction: Maps descriptions like 'labored breathing' to the correct MULTI_SELECT breathing pattern values. Handles variations like 'breath is labored', 'with labored breathing', 'breathing is shallow'.

Dyspnea Severity Mapping: Infers dyspnea severity from descriptions when explicit severity is not stated: 'labored breathing' maps to 'Severe', 'short of breath with exertion' maps to 'Mild', 'severe respiratory distress' maps to 'Severe'.

Assistance Field Normalization: Standardizes variations of assistance levels to schema-compliant values: 'requires assistance' to 'Assisted', 'fully dependent' to 'Dependent', 'able to perform independently' to 'Independent'.

STRING Field Cleaning: Removes units from numeric STRING fields (e.g., '150 mLs' to '150' for PO intake) and converts verbose yes/no responses ('No known allergies', 'NKDA') to simple 'No'. Normalizes whitespace and removes trailing punctuation.

Multi-Select Conjunction Parsing: Splits conjunction phrases for MULTI_SELECT fields: 'nausea and vomiting' becomes ['nausea', 'vomiting'], 'numbness and tingling' becomes separate values. Handles various conjunction patterns ('and', 'with', commas).

4.3 Rule-Based Extraction

The rule-based component employs over 400 regular expression patterns organized by clinical domain and value type. These patterns provide high-precision extraction for common clinical values and serve as supplementary extractions for observations missed by the LLM.

4.3.1 Pattern Organization

Vital Sign Patterns (~65 patterns): Target numeric extractions for oxygen saturation, heart rate, respirations, blood pressure, temperature, mean arterial pressure, urine output, emesis volume, bladder scan volume, oxygen flow rate, and fall risk scores. Patterns handle various unit notations and contextual phrasing.

Categorical Patterns (~180 patterns): Target SINGLE_SELECT fields including oxygen delivery devices, dyspnea severity, cognitive status, orientation levels, Glasgow Coma Scale components, bed position, breath sounds, cough descriptions, fall risk levels, nausea/vomiting, edema characteristics, mobility status, and assistance levels.

Multi-Select Patterns (~155 patterns): Target MULTI_SELECT fields including respiratory interventions, breathing patterns, sensory symptoms, delirium symptoms, urine appearance, and abdomen exam findings. These patterns extract individual values that are later combined.

Field	Pattern Example	Matches
O2 Sat	(?:O2 sat|SpO2)[^\d]*(\d{2,3})\s*%	O2 sat 94%
Heart Rate	(?:heart rate|HR)[^\d]*(\d{2,3})	heart rate 88
Dyspnea	severe\s+dyspnea	severe dyspnea
Cognition	alert\s+(?:but|with)\s+confusion	alert but confused
Breathing	breath(?:ing)?\s+(?:is\s+)?labored	breathing is labored

Table 2: Example regex patterns for different observation types.

4.3.2 Contextual Validation

Contextual validation prevents false positive extractions through several mechanisms:

Keyword Proximity: Cardiac extractions require cardiac keywords within 50 characters of the numeric value. Similarly, respiratory extractions require respiratory context. This prevents numeric values from unrelated measurements being incorrectly categorized.

Unit Compatibility: Values with incompatible units are rejected. For example, a heart rate extraction is rejected if followed by volume units

('cc', 'mL'), which would indicate a fluid measurement rather than a pulse rate.

Substring Safeguards: Patterns avoid matching within larger words. For example, 'labored' should not match within 'nonlabored'. This is implemented through word boundary assertions and negative lookbehind patterns.

Negation Detection: A simple negation check filters extractions where the observation appears in negated context. Negation markers include 'no', 'denies', 'without', 'absent', 'negative for', appearing within 30 characters before the term.

4.4 Schema Validation and Merging

Schema validation ensures all extracted observations conform to the SYNUR ontology before final output. The validation process differs by value type:

NUMERIC Validation: Values are extracted using regex, converted to appropriate numeric types (integer if whole number, float otherwise), and validated against physiological ranges as described in Section 4.2.1.

STRING Validation: Values are cleaned by removing extraneous whitespace, normalizing common patterns (verbose yes/no, unit removal from numeric strings), and truncating excessive length. Empty strings after cleaning are rejected.

SINGLE_SELECT Validation: Values are matched against the schema's enumerated options. Exact matching (case-insensitive) is attempted first. If no exact match, fuzzy matching compares lowercased values using substring containment (e.g., 'moderate' matches 'Moderate dyspnea'). This fuzzy matching can introduce false positives when multiple options share substrings, but empirically improves recall more than it hurts precision.

MULTI_SELECT Validation: Input values (as list or single value) are matched individually against enumerated options. Substring matching is applied with safeguards against partial matches (e.g., 'labored' should not match options containing 'nonlabored').

The merging strategy combines LLM and rule-based extractions as follows: (1) LLM extractions form the base result set; (2) rule-based extractions supplement for observation IDs not already extracted by the LLM; (3) a seen-ID set prevents duplicate extractions; (4) LLM results take priority when both methods extract the same field. This approach leverages LLM comprehensiveness while supplementing with high-precision rule-based findings.

5. Experimental Setup

We evaluate our system on the MEDIQA-SYNUR benchmark using the official evaluation script provided by the shared task organizers. The 122 training samples are used exclusively for few-shot example selection; no fine-tuning or parameter updates are performed on any model component.

All development and error analysis reported in this paper is conducted on the 101-sample development set. This set was used for iterative system refinement, including prompt engineering, filter development, and regex pattern creation. Test set results are submitted separately to the shared task evaluation system and are not included in this paper to ensure fair comparison with other participants.

The complete system is implemented in Python 3.x using the Anthropic API for LLM access, standard library regex for pattern matching, and JSON/JSONL for data handling. Processing the full development set requires approximately 15-20 minutes using sequential API calls with default rate limiting.

6. Results

6.1 Overall Performance

Table 3 presents our results on the development set. The system achieves 74.27% precision, 84.32% recall, and 78.98% F1 score. The system generated 1578 predictions against 1390 gold observations, achieving 1108 true positives.

Metric	Value	Count
Precision	74.27%	1108/1578
Recall	84.32%	1108/1390
F1 Score	78.98%	-
True Positives	1108	-
False Positives	364	-
False Negatives	206	-
Gold Observations	1390	-
System Predictions	1578	-

Table 3: Development set results computed using the official MEDIQA-SYNUR evaluation script.

6.2 Error Analysis

Table 4 provides a quantitative breakdown of errors by type. Of 479 total errors, 'Extra' predictions (false positives) account for 273 (57.0%), followed by 'Missing' observations (false negatives) at 115 (24.0%), and 'Wrong Value' errors at 91 (19.0%).

Error Type	Count	%	Description
Missing (FN)	115	24.0%	Gold obs not extracted
Extra (FP)	273	57.0%	Extracted but not in gold
Wrong Value	91	19.0%	Correct ID, wrong value
Total	479	100%	-

Table 4: Error breakdown by type on development set.

6.3 Field-Level Error Distribution

Table 5 shows the most error-prone fields across all error types. Secondary diagnosis emerges as the most challenging concept, contributing errors across all three categories.

Field	Missing	Extra	Wrong	Total
Secondary diagnosis	6	32	16	54
Gastrointestinal symptoms	0	22	0	22
Cognitive status	6	0	12	18
Patient safety	11	0	0	11
Dyspnea	0	0	10	10

Table 5: Top error-prone fields with breakdown by error type.

The majority of errors (57%) are false positives, dominated by over-extraction of secondary diagnoses (n=32). Despite the speculative language filter, the LLM tends to extract conditions mentioned in clinical context that are not explicitly documented as current diagnoses. For example, a transcript mentioning 'history of diabetes' may trigger extraction of diabetes as a secondary diagnosis when the gold standard only includes active conditions.

Wrong value errors concentrate in cognitive status (n=12) and dyspnea (n=10). Cognitive status errors often involve mapping between similar options such as 'Alert' versus 'Alert and oriented' versus 'Alert with general confusion'. Dyspnea errors reflect difficulty inferring severity levels when not explicitly stated.

Missing observations (n=115) frequently involve patient safety (n=11) and mobility (n=7). These concepts often require multi-step inference from implicit information - for example, inferring patient safety measures from mention of bed alarms or fall precautions without explicit safety documentation.

7. Discussion

7.1 Clinical Utility Assessment

Our system achieves 84.3% recall, suggesting that the majority of clinical observations would be captured in a real deployment scenario. This high recall is valuable for reducing documentation burden, as nurses would not need to manually enter most observations. However, 74.3% precision implies that approximately 1 in 4 extracted observations would require correction.

In practice, this performance profile suggests a semi-automated workflow where the system pre-populates observations for nurse verification rather than fully autonomous documentation. Nurses can quickly confirm correct extractions while correcting errors, potentially providing significant time savings compared to manual entry while maintaining documentation accuracy.

The error distribution has implications for clinical deployment. False positives (over-extraction) are generally easier to catch during review than false negatives (missed observations), as nurses can quickly dismiss incorrect pre-populated values. Missing observations require more careful review against the original dictation. Our system's bias toward recall over precision aligns with this workflow consideration.

7.2 Deployment Considerations

The reliance on a proprietary LLM (Claude Opus 4) presents deployment challenges for healthcare settings. Data privacy regulations (HIPAA in the US, GDPR in Europe) may restrict transmission of patient data to external APIs. While we used the Anthropic API for development, production deployment would likely require: (a) on-premises model hosting, (b) privacy-preserving architectures with data de-identification, or (c) use of open-source alternatives.

Cost considerations are significant for high-volume healthcare documentation. Claude Opus 4 pricing at the time of development was approximately $15/1M input tokens and $75/1M output tokens. Processing the 101-sample development set cost approximately $8-10 USD. For a hospital processing thousands of nursing assessments daily, API costs could be substantial. Smaller, faster models (Claude Haiku, GPT-4o-mini) may offer better cost-performance tradeoffs.

Our rule-based component provides a potential fallback when LLM access is unavailable or cost-prohibitive. While we did not evaluate standalone rule-based performance, the 400+ patterns cover the majority of common vital signs and categorical observations. A tiered approach using rule-based extraction for common cases and LLM for complex narratives could optimize cost-effectiveness.

8. Limitations

Our study has several limitations that should be considered when interpreting results and planning future work:

Synthetic Data: We evaluate exclusively on synthetic nursing dictations generated through a simulation pipeline. Real-world transcripts may exhibit greater linguistic variability, speech recognition errors, background noise artifacts, and documentation inconsistencies. The generalization of our approach to authentic clinical settings remains to be validated.

Proprietary Model: Our use of Claude Opus 4, a proprietary closed-source model, limits reproducibility and raises concerns about deployment in regulated healthcare environments. The model's behavior may change with API updates, and there is no guarantee of long-term availability or consistent pricing.

Development Set Evaluation: All reported results are on the development set, which was used for iterative system refinement including prompt engineering, filter development, and regex pattern creation. This introduces potential overfitting to development-specific patterns. Test set performance, evaluated separately through the shared task submission system, provides a more rigorous assessment of generalization.

No Ablation Studies: We do not report ablation studies isolating the contribution of individual components (LLM-only vs. post-processing vs. rule-based extraction). Such studies would better quantify the value added by each pipeline stage and guide future optimization efforts.

Independent Extraction: Our approach extracts observations independently without modeling inter-observation dependencies. Clinical findings often correlate (e.g., tachycardia with fever, edema with heart failure), and a more holistic approach might improve both extraction accuracy and clinical validity checking.

Single LLM Evaluation: We evaluate only Claude Opus 4 without comparison to other LLMs (GPT-4, Llama, Mistral) or domain-specific models (BioBERT, ClinicalBERT). Comparative evaluation would inform model selection for different deployment scenarios.

9. Conclusion

We presented a hybrid system for schema-guided clinical observation extraction from nursing dictations, combining LLM-based extraction with domain-specific post-processing and rule-based augmentation. On the MEDIQA-SYNUR development set, our approach achieves 78.98% F1 score with 74.27% precision and 84.32% recall.

Our error analysis reveals that secondary diagnosis over-extraction and cognitive status value mapping are primary challenges. The high recall suggests the system could reduce documentation burden in a semi-automated workflow, though precision limitations necessitate human review. We provide detailed methodology documentation to enable reproducibility.

While our results demonstrate the potential of LLMs for clinical documentation automation, practical deployment requires addressing cost, privacy, and validation concerns. Future work should include ablation studies, evaluation on authentic clinical data, comparison across LLM architectures, and exploration of open-source alternatives for deployable solutions.

10. Reproducibility

To facilitate reproducibility, we document key implementation details:

Model Configuration: Claude Opus 4 (claude-opus-4-5-20251101), temperature=0.7, max_tokens=4096. API accessed via anthropic Python package version 0.40+.

Prompt Structure: System prompt includes complete schema (193 concepts with IDs, names, types, enums), 3 few-shot examples from training set, extraction instructions emphasizing explicit-only extraction, and domain-specific guidelines for respiratory, cognitive, and diagnosis fields.

Post-Processing: Six filters (speculative diagnosis, mental status, unit orphans, physiological range, context-based, emesis context) and five corrections (breathing pattern, dyspnea severity, assistance normalization, STRING cleaning, multi-select parsing).

Rule-Based Patterns: 400+ regex patterns organized by domain (vital signs, categorical, multi-select) with contextual validation (keyword proximity, unit compatibility, substring safeguards, negation detection).

References

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. (2022). Large language models are few-shot clinical information extractors. In Proceedings of EMNLP 2022, pages 1998-2022.

Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jin, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical BERT embeddings. In Proceedings of the Clinical NLP Workshop, pages 72-78.

Aronson, A.R. and Lang, F.M. (2010). An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3):229-236.

Baumann, L.A., Baker, J., and Elshaug, A.G. (2018). The impact of electronic health record systems on clinical documentation times: A systematic review. Health Affairs, 37(4):655-663.

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877-1901.

Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., and Buchanan, B.G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics, 34(5):301-310.

Chen, Z., Chen, W., Smiley, C., Shah, S., Borber, I., and Chang, S. (2021). Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In Proceedings of NAACL 2021, pages 1124-1134.

Gu, Y., Tinn, R., Cheng, H., et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1):1-23.

Harkema, H., Dowling, J.N., Thornblade, T., and Chapman, W.W. (2009). ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics, 42(5):839-851.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1-38.

Joukes, E., de Keizer, N.F., de Bruijne, M.C., Abu-Hanna, A., and Cornet, R. (2018). Impact of electronic versus paper-based recording on the quality of clinical documentation in nursing. Studies in Health Technology and Informatics, 247:261-265.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234-1240.

Michalopoulos, G., Corbeil, J.P., Bader, C., Bodenstab, N., and Ben Abacha, A. (2026). Overview of the MEDIQA-SYNUR 2026 Shared Task on Synthetic Nursing Observation Extraction. In Proceedings of ClinicalNLP @ LREC 2026.

Rastogi, A., Zang, X., Srivastava, S., Guez, R., and Dilek, Z. (2020). Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In Proceedings of AAAI 2020, pages 8689-8696.

Savova, G.K., Masanz, J.J., Ogren, P.V., et al. (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.

Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. Nature, 620:172-180.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, volume 30, pages 5998-6008.

Wang, Y., Wang, L., Rastegar-Mojarad, M., et al. (2018). Clinical information extraction applications: A literature review. Journal of Biomedical Informatics, 77:34-49.

MedAware at MEDIQA-EVAL 2026: Vision-Language Model Fine-Tuning with Logprob-Based Score Calibration for Medical Response Evaluation

Ziqi Hao[1,*], **Pengbo Liu**[2,*]
[1]McGill University, [2]Unaffiliated
[1]3801 Rue University, Montreal, QC H3A 2B4, Canada
[1]ziqi.hao@mail.mcgill.ca, [2]liupengbo.work@gmail.com
*Corresponding authors

Abstract

We present MedAware, our MEDIQA-EVAL 2026 system for predicting human ratings of medical QA responses from text and images. We fine-tune Qwen3-VL models (4B/8B/32B) with supervised fine-tuning (SFT), and study GRPO as an optional second stage under both LoRA and full-parameter settings. To handle severe label skew and unstable correlation metrics, we use logprob-based continuous scoring with quantile calibration, converting token probabilities into calibrated metric scores without retraining. This reduces prediction collapse on skewed dimensions and improves metric stability in both English and Chinese. The approach follows the official reference-based shared-task setup and is designed to produce meaningful metric estimates even under extreme class imbalance. In the official shared-task submission setting (8B-LoRA SFT with discrete scoring), our system ranked 3rd on English and 1st among participants on Chinese. Separately, in post-competition offline re-evaluations with logprob scoring, the best tested configuration reaches 0.449 EN-ALL and 0.308 ZH-ALL, while SFT initialization remains critical for effective GRPO.

Keywords: medical evaluation, vision-language model, fine-tuning, score calibration, MEDIQA

1. Introduction

Automated evaluation of medical question answering (QA) systems is essential for scaling quality assessment beyond manual expert review. Unlike general-domain text evaluation, medical response assessment requires judging factual accuracy against clinical evidence, completeness of diagnostic reasoning, and appropriateness of language — often with associated clinical images. The MEDIQA-EVAL 2026 shared task (Ben Abacha and Yim, 2026; Yim et al., 2025a) formalizes this problem: given a medical case, a candidate system response, and gold-standard references, predict human expert ratings across multiple quality dimensions.

Recent work has demonstrated that large language models (LLMs) can serve as effective evaluators of text quality. Zheng et al. (2023) showed that GPT-4 judgments correlate strongly with human preferences on open-ended generation, and Liu et al. (2023) improved evaluation quality by leveraging chain-of-thought reasoning and output token probabilities. However, directly applying these LLM-as-judge approaches to structured medical evaluation presents three challenges: (1) the evaluator must understand domain-specific clinical content and integrate multimodal information (images, text); (2) the task requires simultaneous prediction across multiple quality dimensions with heterogeneous scales; and (3) several evaluation metrics exhibit highly skewed ground-truth distributions — for example, 96.7% of Chinese `writing-style` ratings are 1.0 — causing models to collapse to constant predictions with undefined correlation.

As team MedAware, we address these challenges through fine-tuned Qwen3-VL (Bai et al., 2025) vision-language models with a post-hoc calibration strategy. Our contributions are:

- A two-stage training formulation for multimodal medical evaluation, where supervised fine-tuning (SFT) provides a base evaluator and reinforcement learning (GRPO) is explored as an optional second-stage alignment method.
- A stage-aware comparison across model scales (4B, 8B, 32B): SFT-only, SFT→GRPO, and GRPO-only, while treating parameterization (LoRA vs. full fine-tuning) as an orthogonal choice within each stage. We report primary results on the official test set and use cross-validation for model selection.
- A logprob-based continuous scoring method with quantile calibration that addresses skewed label distributions and undefined correlations, yielding more stable and informative metric estimates.
- Empirical GRPO findings that SFT initialization is critical for GRPO to approach SFT performance and for logprob scoring to transfer effectively.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 192–199
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Related Work

LLM-based evaluation. The paradigm of using LLMs as evaluators has gained significant traction. Zheng et al. (2023) proposed using GPT-4 as a judge for open-ended generation, showing strong correlation with human preferences. Liu et al. (2023) introduced G-Eval, which leverages chain-of-thought reasoning and token probabilities to improve evaluation quality. Our work extends this line by fine-tuning a vision-language model specifically for medical evaluation with structured multi-metric prediction.

Medical NLP shared tasks. The MEDIQA series of shared tasks has driven progress in medical NLP evaluation, including multilingual and multimodal settings.

Reinforcement learning for alignment. Following DeepSeek-R1 (DeepSeek-AI, 2025), GRPO optimizes language models using group-relative advantages without requiring a separate value model. In medical vision-language reasoning, Med-R1 (Chen et al., 2025) further shows GRPO-style training can improve cross-modality and cross-task generalization. For medical image grounding, MedGround-R1 (Xu et al., 2025) adapts GRPO with spatial-semantic rewards and reports strong gains without chain-of-thought supervision. We explore GRPO for aligning model outputs with human evaluation judgments in medical response evaluation.

3. Task Description

The MEDIQA-EVAL 2026 task (Ben Abacha and Yim, 2026; Yim et al., 2025a) requires participants to predict human evaluation ratings for medical question answering systems. Given a medical case (clinical question and context, optionally with images), a candidate system response, and gold-standard reference responses, the system must predict quality ratings on a 3-point scale ($\{0, 0.5, 1.0\}$) across multiple dimensions.

English metrics (6). Each candidate response is rated on a 3-point scale $\{0, 0.5, 1.0\}$ for six dimensions: `disagree_flag` (factual contradiction/error; binary with 1 = disagree and 0 = agree); `completeness` (coverage of question aspects; 1 = complete, 0.5 = partial, 0 = incomplete); `factual-accuracy` (factual correctness against gold references); `relevance` (alignment with the patient question); `writing-style` (appropriateness for patient-facing communication); and `overall` (holistic quality across the above dimensions).

Chinese metrics (2). Chinese cases are rated on: `factual-consistency-wgold` (factual agreement with gold references) and `writing-style` (communication appropriateness). Note that the Chinese `writing-style` distribution is extremely skewed: 96.7% of ratings are 1.0, making correlation-based evaluation particularly challenging.

Evaluation. System performance is measured by the average of Kendall's τ, Pearson's r, and Spearman's ρ correlations between predicted and human ratings, computed per (language, metric) pair and then macro-averaged across all 8 metrics.

Data. The training set contains 161 medical cases (56 IIYI + 105 WoundCare) with 4,872 individual ratings. The IIYI subset is based on DermaVQA (Yim et al., 2024), and the WoundCare subset is based on WoundcareVQA (Yim et al., 2025b). The test set contains 193 cases with ratings from 4 human raters. Both English and Chinese versions are provided for all cases; representative bilingual examples are shown in Figure 1.

Following the official MEDIQA-EVAL 2026 setup, gold-standard reference responses are provided as part of the input at both training and test time. Our method is therefore a *reference-based* evaluator: at inference, the model conditions on the case, candidate response, and provided references, but never on the hidden human ratings used for evaluation.

4. System Description

Our full pipeline (Figure 2) is: model tuning (SFT, GRPO-only, or SFT→GRPO) produces a scorer that generates discrete JSON scores, and post-hoc logprob extraction converts these into continuous scores for calibration.

4.1. Problem Formulation

We formulate evaluation as structured JSON prediction. Given input context x comprising the clinical case (question, context, up to 3 images), a candidate system response, and up to 3 gold-standard reference responses, the model generates a single JSON object containing scores for all relevant metrics in one decoding pass. For English:

```
{"disagree_flag": 0, "completeness": 0.5, "factual-accuracy": 1.0, "relevance": 1.0, "writing-style": 1.0, "overall": 0.5}
```

This joint formulation captures inter-metric dependencies (e.g., low factual accuracy should imply low overall score) without requiring separate inference passes.

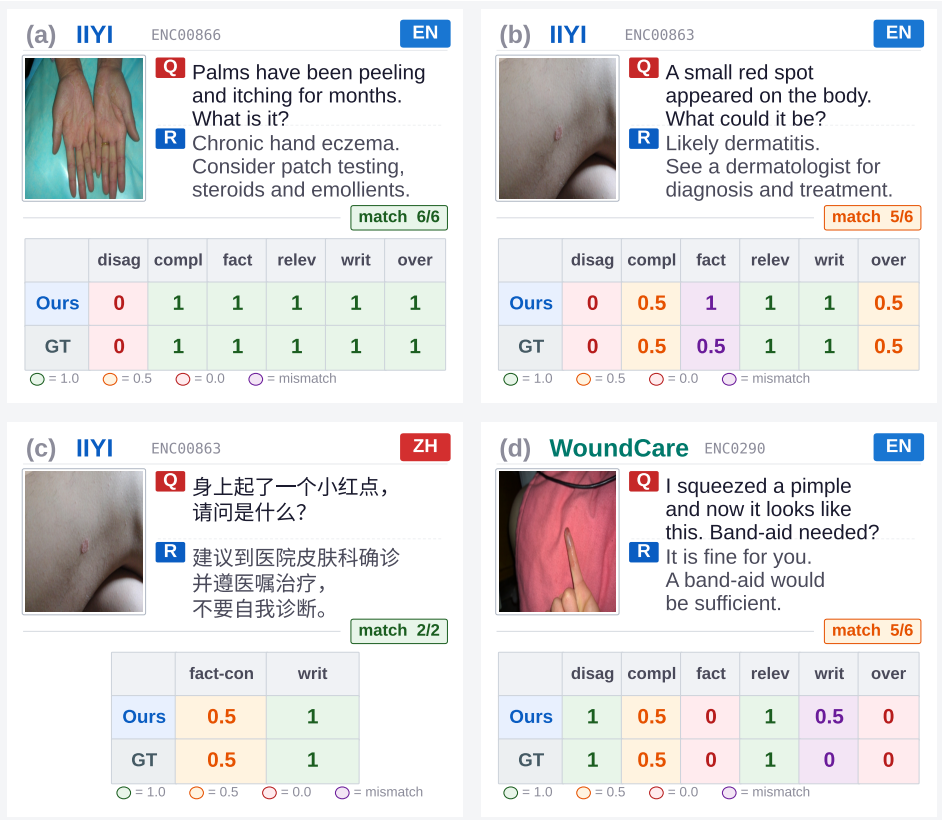


Figure 1: Representative MEDIQA-EVAL 2026 examples from IIYI/WoundCare in English and Chinese. Color badges denote human scores: 1.0, 0.5, 0.0.

Stage 1: Model Tuning (SFT / GRPO)
Input: case + images + candidate + gold refs
Backbone: Qwen3-VL + (LoRA or full FT)
Training mode: SFT / GRPO-only / SFT→GRPO
Output: `{"factual-accuracy": 1.0, "completeness": 0.5, ...}`

Stage 2: Logprob Extraction
At each score token position:
$\hat{s} = P(\text{“1”}) \cdot 1.0 + P(\text{“0”}) \cdot (\pi_{0.5} \cdot 0.5)$

Stage 3: Quantile Calibration
$\hat{s}_{\text{cal}} = F_{\text{GT}}^{-1}(\text{rank}(\hat{s}))$

Figure 2: Pipeline overview of our approach.

4.2. Model and Prompt Design

We use Qwen3-VL (Bai et al., 2025), a vision-language model supporting interleaved image-text inputs via a unified transformer architecture. We experiment with 4B, 8B, and 32B parameter variants.

The input prompt is structured as a chat-style conversation. The **system message** instructs the model to act as a medical expert and provides scoring guidelines (e.g., “Give 0.0 if the system response contradicts the gold standard”). The **user message** concatenates: (1) clinical images (encoded as vision tokens); (2) the patient question; (3) gold-standard reference responses; (4) the candidate response to evaluate; and (5) metric-specific scoring rubrics with an output format template. Separate prompt templates are used for English and Chinese, with language-appropriate metric definitions.

4.3. Supervised Fine-Tuning

We fine-tune using LoRA (Hu et al., 2022) with rank r=16, scaling factor α=32, and dropout 0.05, applied to attention and MLP projection layers. This yields approximately 44M trainable parameters for the 8B variant ($\sim$0.5% of total). We also compare against full fine-tuning for the 4B and 8B models.

Training minimizes cross-entropy loss on the target JSON tokens only, with all input prompt tokens masked (label $= -100$):

$$\mathcal{L}_{\text{SFT}} = -\sum_{t \in \mathcal{T}_{\text{output}}} \log P(y_t \mid y_{<t}, x; \theta) \qquad (1)$$

where $\mathcal{T}_{\text{output}}$ denotes token positions in the target JSON response.

Key hyperparameters: learning rate 2×10^{-5} (linear warmup over 5% of steps), per-device batch size 1 with gradient accumulation over 8 steps (effective batch size 8), AdamW with weight decay 0.01, gradient clipping at 1.0, bfloat16 preci-

sion. We train for 15–30 epochs depending on configuration (30 for 4B/8B LoRA, 20 for 32B-LoRA and 4B-Full, 15 for 8B-Full), evaluating on the held-out fold at every epoch. We select the checkpoint with the lowest validation loss via `load_best_model_at_end`; in practice, the best checkpoints consistently occur at epoch 2–4, with all models overfitting substantially thereafter due to the small training set (130 cases per fold), as shown in Figure 3.

4.4. Group Relative Policy Optimization

We explore GRPO, as applied in DeepSeek-R1 (DeepSeek-AI, 2025), to further align model outputs with human evaluation judgments as a post-competition experiment. We investigate two initialization strategies: (1) starting from an SFT checkpoint (**SFT→GRPO**), and (2) training directly from the pretrained base model (**GRPO only**).

We also run two reward families in exploratory GRPO experiments. The *simple* reward focuses on single-metric accuracy, providing a cleaner gradient signal that is easier to debug and interpret. The *composite* reward combines multi-metric accuracy with small bonuses for exact matches, output format, reasoning tags, and calibration:

$$r_{\text{simple}} = 0.2 \cdot r_{\text{fmt}} + 0.8 \cdot r_{\text{acc-single}}$$

$$r_{\text{composite}} = r_{\text{acc-multi}} + r_{\text{em}} + r_{\text{fmt}} + r_{\text{think}} + r_{\text{cal}}$$

where $r_{\text{fmt}} \in \{0, 1\}$ indicates valid JSON output, $r_{\text{acc-single}} = 1 - |\hat{s} - s^*|$ is distance-based accuracy on one focus metric, and $r_{\text{acc-multi}}$ is the weighted mean of distance-based per-metric accuracies. In the composite reward, $r_{\text{em}}{=}0.5\rho_{\text{exact}}$ uses the per-metric exact-match ratio $\rho_{\text{exact}} \in [0, 1]$, while $r_{\text{fmt}}{=}0.05b_{\text{fmt}}$, $r_{\text{think}}{=}0.05b_{\text{think}}$, and $r_{\text{cal}}{=}0.05b_{\text{cal}}$ depend on valid JSON, the presence of a `<think>...</think>` block, and use of valid score values. All GRPO experiments reported in this paper use thinking mode: the model first produces a brief `<think>...</think>` analysis and then outputs the final JSON. We compare these two reward families as post-competition diagnostics; rollout count K is not treated as an ablation variable in this paper.

GRPO computes per-sample advantages normalized within each group of K candidates:

$$a_j = \frac{r_j - \bar{r}}{\max(\sigma_r, \epsilon)} \tag{2}$$

and updates the policy via a clipped surrogate objective with KL regularization ($\beta{=}0.1$) against the frozen reference model. We use conservative hyperparameters: learning rate $5{\times}10^{-7}$ and max generation length 768 tokens, and train up to 3000 episodes. Unless otherwise stated, all GRPO inference results reported in this paper use checkpoint-1000. At inference time, we apply the same logprob-based scoring and quantile calibration pipeline described in Sections 4.5–4.6.

4.5. Logprob-Based Continuous Scoring

A key challenge is that several metrics have highly skewed ground-truth distributions. For example, Chinese `writing-style` has 96.7% of ratings equal to 1.0. Models trained on discrete targets $\{0, 0.5, 1.0\}$ learn to predict the majority class for all samples, producing zero-variance predictions where correlation is undefined.

We address this by extracting continuous scores from the model's output token probabilities at inference time, requiring no retraining. We run generation with `output_scores=True` to obtain per-position logit vectors. For each metric value in the generated JSON, we locate the corresponding score token position via character-to-token alignment, then compute:

$$\hat{s} = \tilde{P}_1 \cdot 1.0 + \tilde{P}_0 \cdot (\pi_{0.5} \cdot 0.5) \tag{3}$$

where $\tilde{P}_0, \tilde{P}_1$ are the softmax probabilities of tokens "0" and "1" renormalized to sum to 1, and $\pi_{0.5} = P(\text{score}{=}0.5 \mid \text{score}{\neq}1.0)$ is the prior probability of a 0.5 rating among non-1.0 ratings, estimated per metric from the training set (e.g., $\pi_{0.5}{=}0.738$ for `completeness`, 0.0 for binary `disagree_flag`).

We compute $\hat{s}$ using the first digit token ("0" vs. "1") rather than directly using $P(\text{"0.5"})$ because (i) "0.5" is often tokenized into multiple sub-tokens, making probability extraction sensitive to exact token boundaries, and (ii) the first digit token can be aligned more reliably across decoding variants and JSON formatting differences. The class prior $\pi_{0.5}$ then splits the "0" mass between 0.0 and 0.5 in expectation.

The intuition is that even when the model's argmax output is "1" for every sample, the probability mass on "0" varies and captures the model's relative uncertainty. This converts a degenerate constant prediction into a continuous signal with nonzero variance.

4.6. Quantile Calibration

Logprob scores preserve ranking but may not match the ground-truth scale, which matters for Pearson correlation. This is consistent with the broader calibration literature showing that modern neural networks can be miscalibrated (Guo et al., 2017). We apply per-metric quantile calibration: each prediction is mapped to its percentile rank among all predictions for that metric, then mapped to the corresponding quantile of the training set's

ground-truth distribution:

$$\hat{s}_{\text{cal}} = F_{\text{GT}}^{-1}(\text{rank}(\hat{s})) \quad (4)$$

where F_{GT}^{-1} is the empirical quantile function of human ratings. This aligns the marginal distribution of predictions with human ratings while preserving the prediction ranking.

Importantly, calibration uses only training-set labels to estimate F_{GT}; on validation/test splits it uses only the model predictions to compute rank($\hat{s}$) (no access to evaluation labels).

5. Experimental Setup

5.1. Cross-Validation

We perform 5-fold cross-validation on the training set (161 cases) by splitting at the encounter/case level and stratifying folds by dataset (IIYI vs. WoundCare) to ensure balanced representation. Each fold trains on 80% of cases and evaluates on the held-out 20%. This allows us to compare model configurations without touching the test set.

5.2. Configurations

We evaluate five configurations: **4B-LoRA**, **8B-LoRA**, and **32B-LoRA** (all with LoRA rank 16), plus **4B-Full** and **8B-Full** (full fine-tuning). Training uses bfloat16 on NVIDIA H100 80GB GPUs, with 1–4 GPUs depending on configuration and training stage.

6. Results

6.1. Cross-Validation Results

As shown in Table 1, 32B-LoRA attains the highest EN (0.481) and ZH (0.374) under this CV protocol, and is highest on each English metric in this table. Among metrics, `factual-accuracy` is the easiest (∼0.49–0.52) while `writing-style` is the hardest (∼0.31–0.39) for English. Chinese `writing-style` is extremely difficult (∼0.14–0.17) due to distributional skew (96.7% of ground truth ratings are 1.0). LoRA and full fine-tuning perform comparably in CV: full fine-tuning is slightly better at 4B (EN 0.439 vs. 0.435) but worse at 8B on ZH (0.347 vs. 0.360), suggesting that neither consistently dominates on this small dataset.

Compared with discrete scoring, logprob-based scoring mitigates prediction collapse on highly skewed metrics (e.g., Chinese `writing-style`), yielding a continuous signal with nonzero variance and thus well-defined correlation estimates. Quantile calibration further aligns the prediction scale with the training-set rating distribution while preserving ranking.

6.2. Test Set Results

Table 2 reports post-competition offline test-set logprob results for the five SFT configurations and the GRPO models. Table 2 includes only composite-reward GRPO rows; simple-reward 8B diagnostics appear in Table 3. Our official submission was 8B-LoRA with discrete scoring, ranking 3rd on English and 1st among participants on Chinese. These offline test-set numbers should therefore be interpreted separately from the official leaderboard results. Model rankings shift between CV and test: 32B-LoRA leads CV (EN 0.481), while 4B-Full leads test (EN 0.449), consistent with high variance under the small training set (161 cases).

6.3. GRPO Results (Post-Competition)

As shown in Table 2, SFT→GRPO (LoRA) remains close to SFT performance with logprob scoring: EN is 0.428 (4B) and 0.431 (8B), versus 0.437 and 0.433 for the corresponding SFT-LoRA models. Per-metric behavior is similar, with `factual-accuracy` strongest (0.542–0.544) and `completeness` weakest (0.309–0.322). In contrast, GRPO-only settings (LoRA and Full FT) underperform their matched SFT baselines across scales, with EN drops of 0.072–0.111 and `overall` drops of 0.071–0.127.

6.4. Additional GRPO Diagnostics

Table 3 shows that initialization matters more than reward family: within each reward family, SFT→GRPO clearly outperforms GRPO only (EN-logprob gain +0.073–0.074). Figure 4 plots the first 1000 training steps for direct comparison.

LoRA vs. full fine-tuning under GRPO. With composite-reward logprob scoring, GRPO-only LoRA is slightly better on EN at both scales and on ZH at 4B, while Full FT is slightly better on ZH at 8B.

Logprob vs. discrete scoring. Table 4 compares both scoring methods for all five SFT models. Logprob scoring consistently improves EN-ALL, with the largest gains for full fine-tuned models.

For Chinese evaluation, discrete scoring yields undefined ZH-ALL for 3 of 5 SFT models because `writing-style` collapses to a constant prediction. Logprob scoring resolves this and improves EN for all five SFT models; gains are smaller for GRPO (+0.015–0.034 EN). This makes logprob scoring not merely a performance improvement but a practical requirement for evaluating highly skewed metrics with meaningful correlation estimates.

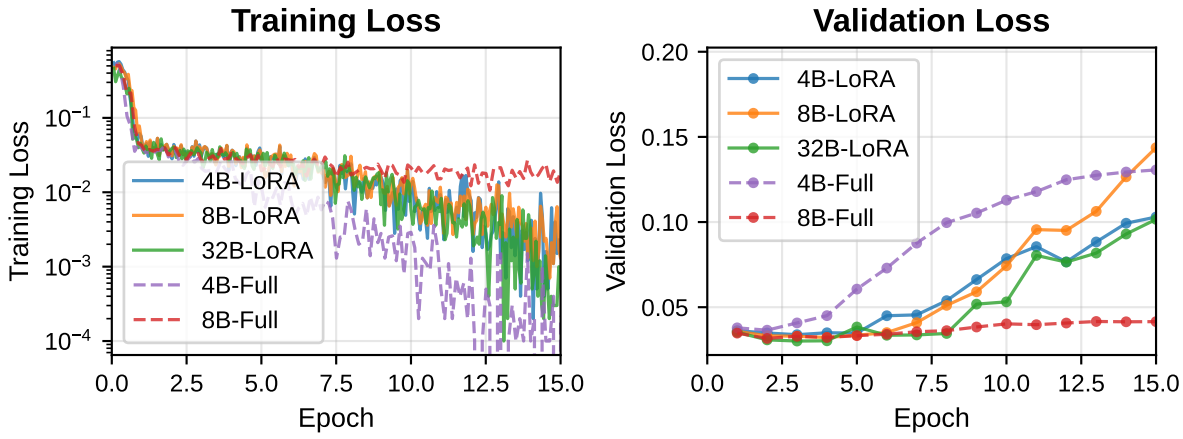


Figure 3: Training and validation loss curves for fold-0 across all configurations. All models overfit after epoch 5–10, with full fine-tuning (dashed) overfitting more severely than LoRA (solid). 32B-LoRA shows the slowest validation loss increase.

	English Metrics							Chinese		
Model	Disag.	Compl.	Fact.	Relev.	Writ.	Over.	**EN**	Fact.	Writ.	**ZH**
4B-LoRA	.429	.406	.490	.463	.333	.486	.435	.565	.169	.367
8B-LoRA	.411	.380	.485	.436	.310	.489	.418	.571	.149	.360
32B-LoRA	**.492**	**.422**	**.517**	**.524**	**.391**	**.542**	**.481**	.578	**.170**	**.374**
4B-Full	.410	.419	.467	.488	.380	.469	.439	**.585**	.156	.370
8B-Full	.387	.389	.499	.450	.355	.474	.426	.558	.137	.347

Table 1: 5-fold CV results with logprob scoring. Each cell is the mean across folds of the averaged (τ, r, ρ) score. EN = average of 6 English metrics; ZH = average of 2 Chinese metrics. Disag. = disagree_flag, Compl. = completeness, Fact. = factual-accuracy/consistency, Relev. = relevance, Writ. = writing-style, Over. = overall.

	English Metrics							Chinese		
Model	Disag.	Compl.	Fact.	Relev.	Writ.	Over.	**EN**	Fact.	Writ.	**ZH**
SFT (trained on CV fold-0, 80% of data):										
4B-LoRA	.385	.336	**.547**	.487	.352	.518	.437	.458	.034	.246
8B-LoRA†	.391	.321	.540	.461	.359	.524	.433	.484	.038	.261
32B-LoRA	**.430**	**.362**	.524	.461	.374	.506	.443	.475	**.100**	.288
4B-Full	.392	.338	.532	**.514**	**.412**	.508	**.449**	**.527**	.088	**.308**
8B-Full	.384	.297	.533	.440	.361	.502	.420	.482	.040	.261
SFT→GRPO (LoRA):										
4B	.387	.322	.542	.472	.337	.510	.428	.457	.043	.250
8B	.399	.309	.544	.465	.336	**.533**	.431	.463	.047	.255
GRPO only (LoRA):										
4B	.345	.239	.444	.371	.260	.406	.344	.323	.063	.193
8B	.376	.227	.509	.340	.244	.444	.357	.383	.025	.204
GRPO only (Full FT):										
4B	.345	.239	.431	.386	.247	.381	.338	.313	.059	.186
8B	.376	.215	.502	.319	.245	.431	.348	.386	.034	.210

Table 2: Test set results (193 cases) with logprob scoring. SFT models use CV fold-0 training data; GRPO rows use composite reward with $K{=}8$. † marks the official submission model; the numbers shown are post-hoc logprob re-evaluations. Red/blue denote best/second-best per column.

6.5. Analysis

Per-metric patterns. `factual-accuracy` is the strongest EN metric across all models (0.43–

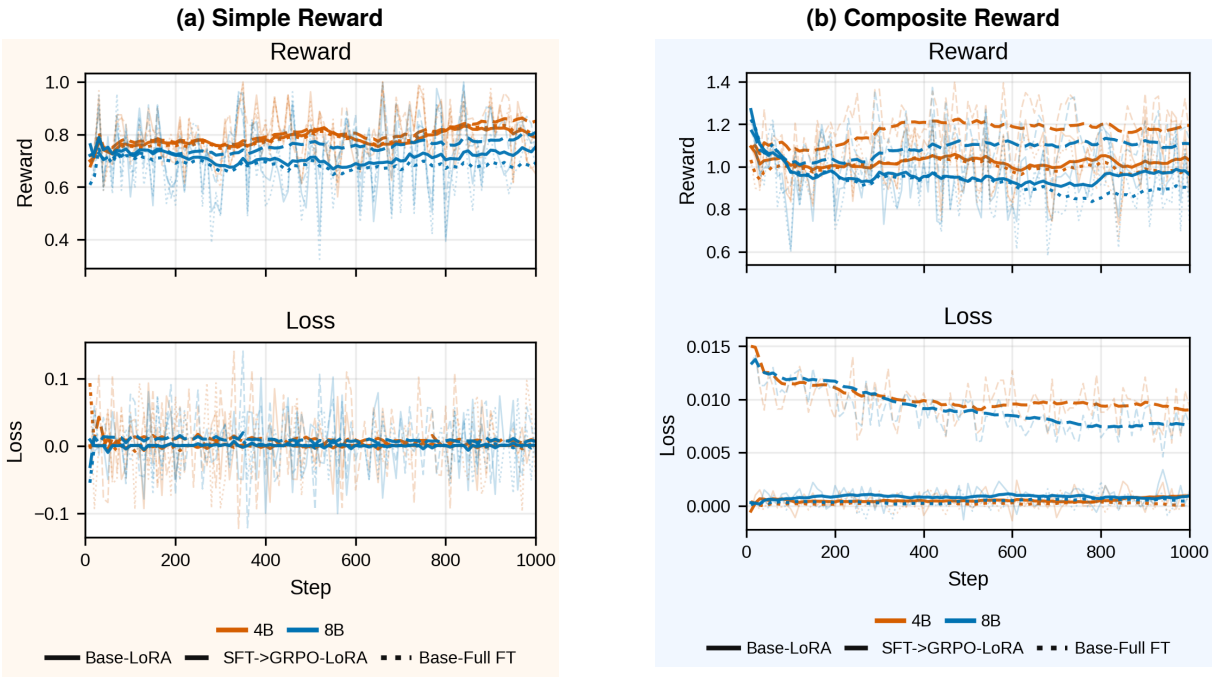


Figure 4: GRPO training curves over the first 1000 steps for simple and composite rewards. Colors denote model scale; line styles denote training strategy (GRPO only LoRA, SFT→GRPO LoRA, GRPO only Full FT).

		Discrete		Logprob	
Reward	**Init**	**EN**	**ZH**	**EN**	**ZH**
Simple	GRPO only	.338	.218	.355	.206
Simple	SFT→GRPO	**.405**	.231	.428	.252
Composite	GRPO only	.341	.218	.357	.204
Composite	SFT→GRPO	.397	**.236**	**.431**	**.255**

Table 3: 8B LoRA GRPO diagnostics comparing reward families and initialization strategies. The composite SFT→GRPO logprob row is the same run as the 8B SFT→GRPO row in Table 2.

	EN			ZH	
Model	Disc.	Logp.	Δ	Disc.	Logp.
4B-LoRA	.399	.437	+.038	—*	.246
8B-LoRA†	.404	.433	+.029	.254	.261
32B-LoRA	.395	.443	+.048	.308	.288
4B-Full	.370	.449	+.079	—*	.308
8B-Full	.296	.420	+.124	—*	.261

Table 4: Discrete vs. logprob scoring on the test set (SFT models). The Logp. columns correspond to the EN/ZH columns in Table 2. †Official submission. *ZH-ALL is undefined because discrete predictions for Chinese `writing-style` collapse to a constant value.

0.55), consistent with CV findings. `completeness` is consistently challenging (0.22–0.36). Chinese `writing-style` is near zero on test (0.03–0.10), far below CV values (0.14–0.17), suggesting poor generalization for this extremely skewed metric.

Model scale and configuration. Under this test split, 4B-Full is the best SFT model (EN 0.449, ZH 0.308), while 32B-LoRA is the strongest LoRA model (EN 0.443). Full fine-tuning also preserves richer logit distributions: 4B-Full is higher than 4B-LoRA on `relevance` (0.514 vs. 0.487) and `writing-style` (0.412 vs. 0.352) with logprob scoring.

CV-to-test gap. The ranking shift from 32B-LoRA in CV to 4B-Full on test suggests substantial fold variance with only 161 cases, making CV unreliable for separating closely-performing configurations. It also suggests that with limited medical evaluation data, smaller fully fine-tuned models may generalize more robustly than larger LoRA-adapted ones.

GRPO analysis. SFT→GRPO preserves the logit structure needed for effective logprob scoring but does not improve on SFT. The likely causes are the small dataset and the low-entropy score space $\{0, 0.5, 1.0\}$, which often collapses reward variance within a GRPO group. GRPO only fails more severely because the pretrained model lacks the structured JSON behavior that both the reward function and logprob extraction rely on.

7. Conclusion

We presented MedAware, a Qwen3-VL-based system for medical response evaluation. Our official submission used 8B-LoRA SFT, ranking 3rd on English and 1st among participants on Chinese. The main finding is that logprob-based continuous scoring with quantile calibration is essential under skewed rating distributions, turning unstable or undefined discrete correlations into usable evaluation signals. For GRPO, the clearest result is the gap between initialization strategies: SFT→GRPO remains competitive, whereas GRPO-only substantially underperforms, indicating that supervised structured generation is a prerequisite for effective RL in this setting.

8. Limitations

Our approach has several limitations. The dataset is small (161 cases), which limits generalization and makes correlation-based comparisons unstable. The evaluator is reference-based, which matches the shared-task setup but limits direct use in settings without expert reference responses. Logprob scoring relies on priors estimated from the training distribution and may transfer poorly under distribution shift. GRPO was explored only in a limited regime, and longer training, larger-scale runs, and stronger cross-metric calibration remain future work.

9. Acknowledgements

We thank the MEDIQA-EVAL 2026 organizers for creating this shared task and providing the dataset.

10. Bibliographical References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Asma Ben Abacha and Wen-wai Yim. 2026. Overview of the MEDIQA-EVAL 2026 shared task on evaluation metrics in medical multimodal question answering. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

Haotian Chen, Han Ding, Yiru Bai, Ren Gao, Junbo Yang, Yuheng Li, and Konstantinos Psounis. 2025. Med-R1: Reinforcing medical reasoning in MLLMs via GRPO with thought preferences. *arXiv preprint arXiv:2503.13939*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Huihui Xu, Yuanpeng Nie, Hualiang Wang, et al. 2025. MedGround-R1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15964, pages 391–401. Springer Nature Switzerland.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2025a. MORQA: benchmarking evaluation metrics for medical open-ended question answering. *CoRR*, abs/2509.12405.

Wen-Wai Yim, Asma Ben Abacha, Robert Doerning, et al. 2025b. Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care. *Journal of Biomedical Informatics*, 170:104888.

Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2024. DermaVQA: A multilingual visual question answering dataset for dermatology. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005, pages 209–219. Springer Nature Switzerland.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

HSE NLP TEAM at MEDIQA-SYNUR 2026: Consensus Adjudication Ensemble (ACE): Balancing Precision and Recall for Schema-Bystander Clinical Extraction

Airat Valiev
HSE University
Moscow, Russia
aa.valiev@hse.ru

Abstract

Clinical documentation from nurse dictations is labor-intensive and error-prone, yet it contains high-value observations that must be transferred into structured flowsheets. The MEDIQA-SYNUR 2026 shared task evaluates systems that extract and ontology-align 193 clinical concepts (with heterogeneous value types) from synthetic speech transcripts derived from intensive care notes. We describe the Consensus Adjudication Ensemble (ACE), a three-stage pipeline that (i) maximizes candidate coverage via complementary generators, (ii) enforces high precision through a dedicated adjudicator that operates as a verifier rather than a generator, and (iii) restores strict schema compliance using a targeted, token-efficient repair step. On the official test set we achieve an exact-match micro-F1 of 0.7996 (P=0.7812, R=0.8188), ranking 4th on the leaderboard. Beyond the competitive result, we analyze clinically relevant failure modes - hallucinated interventions, over-confident categorical labels, and unit/normalization errors - and quantify adjudication trade-offs: 2,219 candidates removed, 91.3% of which are true false positives, at the cost of 8.7% mistakenly removed true positives. Finally, targeted schema repair reduces validation context from $\sim$230k tokens to $<$2k per document while preserving most extraction gains.

Keywords: clinical information extraction, nurse dictations, ontology alignment, ensemble methods, adjudication, error analysis

1. Introduction

Clinical flowsheets are the substrate of routine inpatient care: they record vital signs, symptoms, assessments, and interventions at high temporal resolution. In many workflows, these fields are populated from speech dictations and later normalized into discrete slots, which introduces delays and errors. Documentation burden is substantial - often reported as a large fraction of clinical workload - and motivates automation of structured capture from free-form dictation.(Panchal and Thakur, 2024; Moy et al., 2023; Mitha et al., 2023) The MEDIQA-SYNUR 2026 shared task targets this bottleneck by requiring systems to extract structured clinical observations from synthetic nurse dictations and align them to a fixed ontology.(Michalopoulos et al., 2026)

We focus on a high-stakes regime where *precision is safety-critical* (hallucinated interventions and symptoms are unacceptable), while *recall impacts completeness* (missed observations degrade downstream decision-making). Contemporary LLM-based extractors tend to entangle these objectives; pushing recall typically increases hallucination. Our solution, the **Consensus Adjudication Ensemble (ACE)**, decouples them into separate stages: (1) generate a diverse candidate pool, (2) adjudicate as a verifier, and (3) repair schema violations cheaply.

This paper makes three contributions: (i) A practical **generation - verification decomposition** for ontology-aligned extraction from noisy speech transcripts, implemented as an ensemble with a dedicated adjudicator; (ii) A **targeted schema repair** mechanism that validates and fixes only the subset of outputs that violate constraints, reducing validation context by two orders of magnitude; (iii) A **granular error analysis** (false positives, false negatives, and value errors) and ablations that expose when retrieval augmentation helps and when it induces "contextual copying" hallucinations.

2. Related Work

LLMs for clinical information extraction. Clinical entity and attribute extraction has historically relied on supervised sequence labeling and domain-adapted encoders, which provide high precision but require task-specific training and struggle with rapidly changing schemas.(Alsentzer et al., 2019) Instruction-tuned LLMs enable flexible, schema-conditioned extraction without training data, including few-shot clinical extraction settings.(Agrawal et al., 2022) However, they often violate strict biomedical constraints (enumerations, units, and type consistency) unless controlled by explicit validation or constrained decoding.

Consensus and verification. Ensembling and verification-style decomposition have been used to improve robustness of generative systems. For

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 200–211
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Split	Ex.	Avg tok.	Avg obs.	Uniq.
Train	122	275.3	13.8	166
Dev	101	279.9	13.0	170
Test	199	259.3	12.8	164

Table 1: SYNUR dataset statistics by split. Value-type proportions are reported in Table 2.

Split	%SS	%MS	%Num	%Str
Train	71.2	11.6	10.8	6.4
Dev	72.3	11.9	10.4	5.3
Test	72.6	13.5	10.1	3.8

Table 2: Value-type proportions by split (SS/MS/Num/Str denote SINGLE_SELECT, MULTI_SELECT, NUMERIC, STRING).

extraction, the key benefit is role separation: candidate generation can optimize recall, while a verifier (adjudicator) can optimize precision by requiring explicit grounding. Related ideas include self-consistency aggregation for reasoning,(Wang et al., 2023) multi-agent ensembling,(Li et al., 2024) and LLM-as-a-judge paradigms where a strong model evaluates and filters outputs of other systems.(Zheng et al., 2024) This is especially important in clinical documentation, where plausible-but-unstated facts are harmful.

Risks of retrieval augmentation. RAG is effective for knowledge-intensive QA, but for schema-conditioned extraction it can be counterproductive: retrieved exemplars may leak surface forms and induce spurious predictions - LLMs are distracted by irrelevant context (Shi et al., 2023) and exhibit position-sensitive failures under long contexts (Liu et al., 2024). We empirically observe a strong small-model pathology (contextual copying), motivating safeguards such as strict adjudication and limited retrieval exposure.

3. Task

MEDIQA-SYNUR 2026 evaluates systems that map each dictation transcript to a set of $(\texttt{concept_id}, \texttt{value})$ pairs from an ontology of 193 concepts.(Michalopoulos et al., 2026) Concepts have heterogeneous value types: `SINGLE_SELECT`, `MULTI_SELECT`, `STRING`, and `NUMERIC`. The official metric is micro-averaged exact-match F_1 over concept - value pairs. The dataset is synthetically generated from clinical notes to emulate dictation style while preserving privacy.(Corbeil et al., 2025)

3.1. Dataset statistics (official splits)

To anchor subsequent analyses, Table 1 reports split-level statistics transcribed from our internal draft artifacts. Test labels were released after the competition, so all error analyses in this paper are post-competition.

3.2. Evaluation protocol and strictness of matching

Submissions are scored with micro-averaged precision/recall/F_1 over exact-match $(\texttt{concept_id}, \texttt{value})$ pairs as defined by the organizers.(Michalopoulos et al., 2026) Two details are crucial in practice: (i) enum-valued concepts (`SINGLE_SELECT`/`MULTI_SELECT`) require *membership* in a fixed allowed list (schema compliance is part of correctness), and (ii) any value mismatch counts as fully incorrect, even when the concept is correct. This creates a regime where seemingly "minor" normalization issues (list-vs-scalar, casing, synonym choice) can dominate errors, motivating our explicit repair step.

3.3. Post-competition analyses

The leaderboard ranking is based on the competition submission. Unless stated otherwise, deeper error analyses and cost accounting are performed after the test labels were released and are reported as post-competition experiments.

4. System Overview

ACE is a three-stage pipeline operating on a single transcript x. Let $\mathcal{S}$ denote the schema (concepts, value types, and allowed enumerations). The system produces a set of candidates $C = \{(i, v)\}$ with $i \in \mathcal{S}$ and then filters and repairs them into the final prediction $\hat{Y}$.

4.1. Stage 1: Candidate Generation via Complementarity

We use two generators with different failure profiles and then take their union. In our experiments, a large model with retrieval augmentation improves coverage of infrequent concepts, whereas a smaller/cheaper model without retrieval is less sensitive to distractor context but tends to omit long-tail concepts. The union step is intentionally recall-oriented and accepts that it may be noisy.

Models We deployed two diverse extractors: GPT-5 in a zero-shot configuration,(Singh et al.,

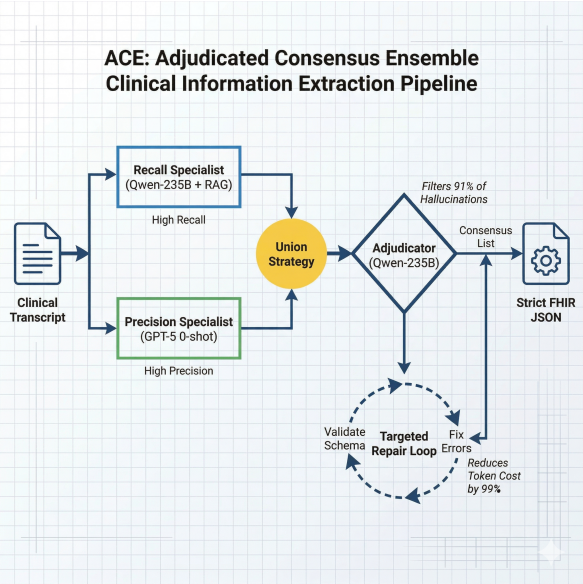


Figure 1: ACE pipeline: high-recall candidate generation (two complementary generators), master adjudication (verification), and targeted schema repair.

2025) and a Qwen3-235B-A22B-Instruct-FP8 (accessed via the Doubleword batch API) with retrieval augmentation(Team, 2025) . Baseline comparisons were also conducted against GPT-4o.(OpenAI, 2024)

4.1.1. Schema-bystander failure mode

Many task concepts are correlated in real clinical notes (e.g., oxygen device co-occurs with respiratory distress), but the dictations in this benchmark may mention only a subset. We found it useful to treat the ontology and any retrieved schema descriptions as *bystanders*: they are necessary for formatting but should not be treated as evidence. This motivates a strict separation between (i) generators that may leverage schema for coverage and (ii) adjudication that requires transcript-grounded support.

4.1.2. Retrieval augmentation and "contextual copying"

RAG can improve recall by making long-tail labels salient, but it also introduces an adversarial surface-form effect: smaller models may copy enumerated values or schema snippets into the output even when the transcript does not support them. In our runs this manifested as spurious categorical labels (e.g., cognitive status) and intervention assertions that look plausible but are ungrounded. We therefore constrain RAG to the high-capacity generator and rely on adjudication to suppress RAG-induced false positives.

Output normalization. Each generator outputs JSON records with a concept id and value. Before union, we normalize trivial formatting differences (e.g., `true/false` vs `Yes/No`, list-vs-scalar for multi-select, and whitespace/casing) to reduce spurious mismatches.

4.2. Stage 2: Master Adjudication (Verification)

Given transcript x and a candidate set C, the adjudicator returns a subset $A \subseteq C$ of candidates that are *entailed* by x under task guidelines. Operationally, the adjudicator is prompted as a verifier: it must *reject* candidates unless there is explicit textual support (or a task-allowed paraphrase) in the dictation.

Relation to LLM-as-a-judge. The master adjudicator follows principles similar to LLM-as-a-Judge,(Zheng et al., 2024) evaluating whether each candidate is supported by the transcript and suppressing clinically plausible but unstated inferences.

4.2.1. Adjudication policy: conservative acceptance criteria

We found that most high-impact hallucinations come from *clinical priors* rather than transcription artifacts (e.g., inferring oxygen therapy from "short of breath"). The adjudicator therefore follows a conservative policy: accept only (i) direct lexical matches (including common paraphrases) or (ii) measurement-backed implications explicitly stated (e.g., "sat 92% on nasal cannula" supports both saturation and oxygen device). Ambiguous hedges ("might", "possibly") are treated as evidence only when the task guidelines permit recording suspected findings.

4.2.2. Cost of verification

Verification is cheaper than regeneration because the adjudicator sees only the candidate set rather than the full schema. This makes it practical to scale the number of candidate generators, which empirically improves recall without paying the full cost of multiple schema-conditioned extractions.

4.3. Stage 3: Targeted Schema Repair

Even after adjudication, outputs may violate schema constraints: wrong enumerations, scalar vs list mismatches, or minor normalization errors. Re-running generation with full schema context is token-expensive. Instead, we perform *targeted repair*: only candidates that fail validation against $\mathcal{S}$ are sent to a repair model together with (i) the

Algorithm 1: Consensus Adjudication Ensemble (ACE)

Input: transcript x, schema $\mathcal{S}$
Output: prediction set $\hat{Y}$

1. $C_1 \leftarrow \textsc{Generate}(x)$ (generator 1)
2. $C_2 \leftarrow \textsc{GenerateRAG}(x)$ (generator 2)
3. $C \leftarrow \textsc{Normalize}(C_1 \cup C_2)$
4. $A \leftarrow \textsc{Adjudicate}(x, C)$
5. $E \leftarrow \{(i, v) \in A : \neg \textsc{Valid}_{\mathcal{S}}(i, v)\}$
6. $R \leftarrow \textsc{Repair}(x, E, \mathcal{S})$
7. $\hat{Y} \leftarrow (A \setminus E) \cup R$
8. **return** $\hat{Y}$

Figure 2: High-level ACE procedure. Repair is applied only to schema-invalid items.

transcript snippet around evidence, (ii) the concept definition and allowed values, and (iii) the invalid output. This reduces validation context from ∼230k tokens to <2k tokens per document in our setup.

Relation to guided generation. To enforce schema compliance without expensive re-sampling, our targeted repair loop is inspired by guided generation, constrained decoding, and structured prompting techniques.(Willard and Louf, 2023; Beurer-Kellner et al., 2023)

4.3.1. Repair taxonomy

In our error logs, repair requests fall into three recurring classes: (1) **boolean/categorical canonicalization** (e.g., `True`→`Yes`); (2) **enum projection** (mapping free-form text to the closest allowed enumeration, e.g., "borderline inadequate"→`inadequate`); (3) **container normalization** (scalar→list for `MULTI_SELECT` concepts). These are low-entropy transformations, which is why they can be handled reliably with small context and without access to the full ontology.

5. Algorithm

6. Experimental Setup

6.1. Evaluation

We report micro-precision, micro-recall, and micro-F_1 on exact match. We additionally compute granular counts of false positives (FP), false negatives (FN), and value errors per concept to support error analysis.

6.2. Granular scoring protocol

To diagnose what drives exact-match failures, we decompose errors into: **FP** (spurious concept - value pair), **FN** (missed gold pair), and **value error** (predicted concept matches gold concept, but value mismatches due to synonymy, formatting, or unit normalization). For `MULTI_SELECT` concepts, we treat each selected value as a separate pair, which makes partial correctness visible as mixed FP/FN rather than a single opaque mismatch.

Statistic	Value
Test examples	199
Avg. tokens/transcript	177.0
Avg. observations/example	12.8
Unique concepts present	164

Table 3: Test-set statistics.

Value type	Count
`SINGLE_SELECT`	1853
`MULTI_SELECT`	344
`NUMERIC`	257
`STRING`	98

Table 4: Value-type distribution on the test set.

6.3. Dataset statistics

From our analysis of the test split (199 examples), transcripts are short (avg. 177 tokens) but dense (avg. 12.8 observations/example), with 164 unique concepts present in the test data. The majority of labels are categorical: `SINGLE_SELECT` dominates, while `STRING` and `NUMERIC` are comparatively rare.

6.4. Precision - recall dynamics across stages

Figure 3 highlights a key design decision: we deliberately tolerate a low-precision union as long as adjudication can remove noise with minimal recall loss. This works well when hallucinations are "obvious" (unsupported interventions, implausible categorical labels) but is less effective when the transcript expresses the observation indirectly or via long-range discourse, in which case the adjudicator may remove true positives.

7. Results

On the official leaderboard we obtain $F_1 = 0.7996$ (P=0.7812, R=0.8188), ranking 4th. Table 5 reports our component-level results. The main qualitative pattern is stable: union substantially increases recall but collapses precision; adjudication recovers precision at a non-trivial recall cost; the final system combines consensus filtering with targeted repair.

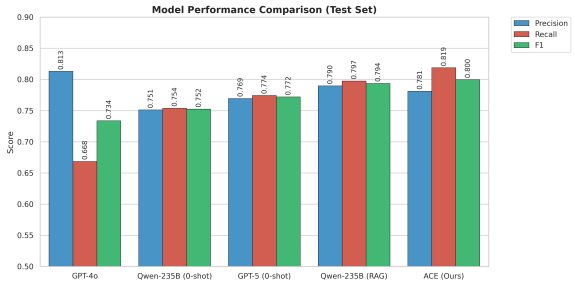


Figure 3: Precision - recall behavior across ACE stages. Union maximizes recall but is noisy; adjudication recovers precision; repair restores schema compliance.

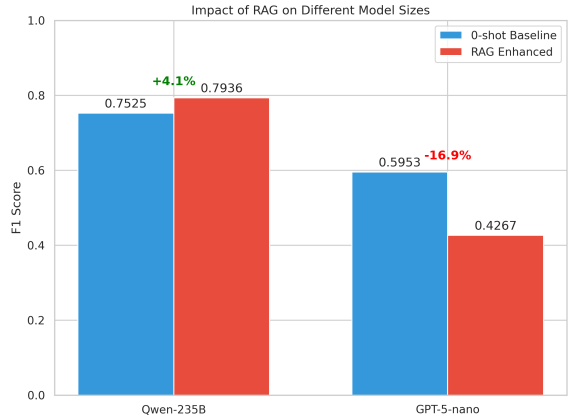


Figure 4: Micro-F_1 across system variants.

7.1. RAG sensitivity across model sizes

RAG had a divergent effect depending on model capacity. For the large model (Qwen-235B), RAG improved F_1 by +4.1 points (0.7525→0.7936) in our experiments; for the smaller model (GPT-5-nano), RAG significantly degraded quality (Table 6).

Contextual copying. In 91% (181/199) of test cases, the small-model RAG variant predicted at least one value absent from the transcript but present verbatim in retrieved examples. We refer to this as *contextual copying* and view it as a primary reason to separate recall-oriented generation from a strict verifier.

8. Efficiency and Cost

Beyond accuracy, we analyze token usage and estimated cost, since practical deployment of large-model pipelines depends on throughput. Table 7 summarizes stage-wise token usage and cost estimates from our internal runs. The key observation is that targeted repair is cheap relative to extraction and thus a favorable place to enforce schema compliance.

Component	P	R	F_1
Ens. union	0.4618	0.7449	0.5702
Adj. consensus	0.7352	0.6582	0.6946
ACE	**0.7812**	**0.8188**	**0.7996**

Table 5: Core component results. "ACE" corresponds to the official submission (ensemble + adjudication + targeted repair). "Ens. union" is the recall base; "Adj. consensus" is the verifier-style filter.

Model setting	P	R	F_1
0-shot	0.7126	0.5112	0.5953
RAG	0.3807	0.4855	0.4267

Table 6: GPT-5-nano RAG effect analysis: retrieval degrades F_1 by 16.9 points due to hallucination/leakage.

9. Error Analysis

We categorize errors into: (i) **FP**: predicted concept - value absent in gold; (ii) **FN**: missed gold item; and (iii) **Value error**: correct concept but mismatched value (often casing, synonyms, list-vs-scalar, or unit normalization).

9.1. Error type distribution

Across the most problematic concepts, we observe a characteristic pattern: high FN counts for implicitly stated activities (e.g., breathing exercises) and high value-error counts for appearance-like multi-select attributes where the ontology enumerations are restrictive (e.g., urine appearance, skin condition). This suggests that improving performance beyond our rank will likely require (i) better discourse-sensitive extraction for implicitly referenced items and (ii) a more principled lexical mapping layer for constrained enumerations.

9.2. Top error types

Table 8 reports the dominant error drivers by type (computed after test labels became available). Respiratory interventions (concept `3`) account for a large share of misses, while urinary symptoms (`177`) are prominent in both hallucinations (FP) and value mismatches, indicating that the ontology constraints are both hard to satisfy and easy to over-predict.

Two mechanisms explain these patterns. First, care actions such as respiratory exercises are often expressed as brief imperatives ("deep breathe", "incentive spirometer") and may be separated from other respiratory descriptors, which increases false negatives under exact-match scoring. Second,

Stage	Model	In tok.	Out tok.	Cost ($)
Extract	Qwen-235B	801,893	33,593	0.09
Extract	GPT-5	503,393	33,593	0.48
Adjud.	Qwen-235B	255,234	33,593	0.04
Repair	Qwen-235B	138,000	3,450	0.02
Total		1,698,520	104,229	0.63

Table 7: Token usage and cost estimation (internal accounting). Stages are extraction, adjudication, and targeted repair; percentages are omitted for compactness.

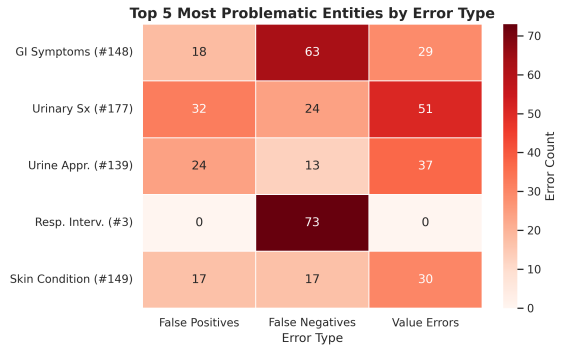


Figure 5: Heatmap of errors by concept. We observe clusters of value normalization errors for multi-select appearance/skin concepts and high FN rates for breathing exercises.

symptom inventories (e.g., urinary symptoms) are high-cardinality multi-select fields; generators tend to over-complete them from clinical priors, and even when the concept is correct, mapping free-form phrasing to the restricted enum set yields value errors.

9.3. Most problematic concepts

From granular error logs, the most frequent *misses and value errors* concentrate in a small subset of concepts (Table 9). Many correspond to implicit or discourse-level information, or to highly variant lexical realizations.

9.4. Hallucination profile and adjudication impact

Adjudication removes 2,219 candidate observations, of which 2,025 (91.3%) are correctly removed false positives. The remaining 194 (8.7%) removed items are true positives, representing a dominant source of recall loss in the ACE pipeline.

Precision - recall impact ratio. This trade-off is strongly favorable for precision: for every true positive removed by the adjudicator, it removes $2025/194 \approx 10.4$ hallucinated observations. We interpret this as a clinically meaningful operating point, because the removed false positives are frequently high-risk categories (fabricated interventions and cognitive-state labels), consistent with self-verification approaches for hallucination detection.(Manakul et al., 2023)

Type	C_ID	Count	Description
FN	3	73	Respiratory interventions
FN	148	63	Gastrointestinal symptoms
Value	177	51	Urinary symptoms (W/A)
FP	177	32	Urinary symptoms (halluc.)

Table 8: Top error types on the test set (post-competition analysis). W/A means Wrong Answer by value, and halluc. is a hallucinated entity.

ID	Concept name (schema)	#FN
148	Gastrointestinal symptoms	92
177	Urinary symptoms	75
3	Respiratory interventions	73
139	Urine appearance	50
149	Skin condition	47
84	Abdomen assessment	39
165	Breath sounds	34
13	Breathing pattern	32
0	Broset violence checklist - confusion	23
130	Cognitive status	20

Table 9: Top 10 most problematic concepts (false negatives + value errors) from the test-set analysis logs, with official schema names.

Where adjudication fails. The dominant failure mode is *indirect evidence*: dictations may describe an observation using a dispersed set of cues (e.g., symptoms spread across narrative) or with hedges and revisions. In such cases, generators may propose a correct item, but the verifier may not find sufficient explicit support to accept it.

Most hallucinated concepts. The same few concepts dominate false positives, often reflecting clinical priors (e.g., inferring comorbidities) or schema copying (multi-select fields). In our logs, `177` (urinary symptoms) and `31` are the top FP drivers, followed by breathing pattern (`13`) and respiratory exercises (`3`).

9.5. Qualitative examples

Adjudication is particularly effective at rejecting clinically plausible but unsupported interventions and cognitive-state labels; repair fixes minor value canonicalization without re-running extraction.

ID	Concept name (schema)	FP
177	Urinary symptoms	32
31	Diabetes history	31
13	Breathing pattern	29
3	Respiratory interventions	25
139	Urine appearance	24
26	Cognitive disturbances	22
148	Gastrointestinal symptoms	18
130	Cognitive status	18
165	Breath sounds	18
0	Broset violence checklist - confusion	18

Table 10: Top 10 most hallucinated concepts (FP) from the test-set analysis logs, with official schema names.

Quantity	Count	Share
Removed by adjudication	2219	100.0%
Correctly removed (FP)	2025	91.3%
Incorrectly removed (TP)	194	8.7%

Table 11: Adjudication effect from our analysis logs.

Adjudication removal. In a representative case, the union stage proposes multiple unsupported items (e.g., cognitive status labels and interventions) that are clinically plausible but not stated; the adjudicator removes them, improving precision.

Repair canonicalization. In separate cases, repair fixes low-entropy mismatches such as boolean-to-enum mapping (`True`→`Yes`), near-miss enum projection ("borderline inadequate"→`inadequate`), and scalar-to-list normalization for `MULTI_SELECT` fields.

10. Implementation Details

We provide implementation details to support replication. All steps operate on a single transcript and an immutable schema $\mathcal{S}$.

10.1. Prediction representation

We represent predictions as a set of JSON objects with keys: `id` (string concept id) and `value`. Values are typed by the concept's `value_type` in the schema. For `MULTI_SELECT`, we normalize values to a JSON list and score each selected value as a separate $(\texttt{id}, \texttt{value})$ pair.

10.2. Schema validation

Before and after adjudication we validate each candidate against $\mathcal{S}$: (i) concept id must exist; (ii) the predicted value must match the declared value type; (iii) for `SINGLE_SELECT`/`MULTI_SELECT`, each value must appear in `value_enum`; (iv) for `NUMERIC`, we require a parseable numeric value and canonicalize trivial formatting (e.g., stripping units when the schema expects a bare number).

Stage	Model	Temp.	Max tok.	Top-p
Extract.A	Qwen3-235B	0.0	2048	1.0
Extract.B	GPT-5	0.0	2048	1.0
Adjudic.	Qwen3-235B	0.0	2048	1.0
Repair	Qwen3-235B	0.0	512	1.0

Table 12: Inference parameters used in each ACE stage.

10.3. Normalization rules

We apply lightweight normalization to reduce avoidable mismatches: **booleans** (`true/false`, `yes/no`) are canonicalized to the schema enum when applicable; **container** normalization maps scalar→list for `MULTI_SELECT`; **casing/whitespace** are normalized for string-like fields. We avoid aggressive synonym mapping at this stage because it can introduce false positives; instead, enum projection is handled by the targeted repair step with explicit access to allowed values.

10.4. Targeted repair prompts (high level)

For each invalid item (i, v), the repair model is given: (a) transcript x (or a short evidence window), (b) concept name and value type, (c) the allowed enumeration if present, and (d) the invalid value. The model must output a single corrected value or an explicit `null` to drop the item. This design bounds context length and makes repair a low-entropy transformation rather than re-extraction.

10.5. Hyperparameters

All stages use greedy decoding (temperature 0.0) to ensure deterministic outputs; parameters are summarized in Table 12.

11. Discussion

11.1. Why verification helps more than better prompting

In this task, many clinically "reasonable" facts are *not stated* in the transcript. Extractor prompts that encourage completeness tend to over-infer. By reframing the second stage as an entailment-style verifier, we obtain a controllable knob for precision without forcing generators to become conservative.

Risk	Error mode	Examples
High	Fabricated interventions / status	Oxygen delivery device; suctioning; "disoriented" cognitive labels
Medium	Omitted symptoms / care actions	Missed GI symptoms; missed respiratory exercises
Low	Normalization / schema mismatch	list-vs-scalar for MULTI_SELECT; enum projection; unit/casing variants

Table 13: Risk-oriented error taxonomy used in our analysis.

11.2. What limits recall after adjudication

The main failure case for a strict verifier is *distributed evidence*: the transcript may mention an observation in a fragmented way (disfluencies, self-corrections, or references like "same as before"), so neither the candidate nor the verifier sees a crisp support span. Future work could use evidence selection (span extraction) as an intermediate representation, allowing the adjudicator to operate on explicit spans rather than raw text.

11.3. Clinical safety lens

We found it useful to interpret errors not only by F_1 but by clinical risk. High-risk FPs include fabricated interventions (oxygen delivery device, suctioning) and misrepresented cognitive status (e.g., disorientation). Medium-risk errors include symptom omissions that may delay escalation. Low-risk errors include unit formatting and mild synonym mismatches. This framing matches why adjudication is valuable: it preferentially removes high-risk hallucinations even when it slightly harms recall.

This taxonomy is supported by our empirical error distributions: high-risk items are over-represented among frequent false positives (Table 10), while medium-risk omissions align with the most frequent false negatives (Table 8 and Table 9). Low-risk issues largely correspond to value and formatting mismatches that are addressed by schema-aware repair.

11.4. Safety implications by model component

We observe a consistent safety pattern across variants: (i) the recall-oriented union stage is the primary source of high-risk hallucinations; (ii) adjudication removes most of these, effectively acting as a safety filter; and (iii) targeted repair mainly reduces low-risk schema mismatches (container and enum canonicalization) without meaningfully affecting the hallucination rate. This supports a design where safety-critical filtering is placed before normalization-heavy post-processing.

11.5. When retrieval augmentation helps (and when it hurts)

We observe that retrieval augmentation is beneficial for large models by improving coverage of long-tail concepts, but can be harmful for smaller models: they exhibit contextual copying behavior, producing surface forms from retrieved schema snippets even when unsupported by the transcript. This motivates keeping RAG in the high-capacity generator and relying on adjudication to suppress spurious candidates.

12. Limitations and Ethics

The dataset is synthetic and derived from ICU notes; dictation style and concept prevalence may differ in other departments or spoken workflows. Verifier-style adjudication can remove true positives when evidence is indirect or distributed across the transcript. Finally, while we emphasize precision to reduce hallucinated observations, our system is not a clinical device and must be validated prospectively before any deployment. Comparisons with additional small and mid-range models (e.g., Qwen3-4B) remain as future work due to competition timeline constraints.

13. Data and Code Availability

Our code is available in an anonymized repository. The repository contains inference scripts and prompt. [1]

14. Bibliographical References

Monica Agrawal, Stefan Hegselmann, Hunter Hunter, 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 1997–2022.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the*

[1] https://anonymous.4open.science/r/MEDIQA-SYNUR-HSE_NLP-8C2B/

2nd Clinical Natural Language Processing Workshop, pages 72–78. Association for Computational Linguistics.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2023. Prompting is programming: A query language for large language models. In *Proceedings of the 44th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 1826–1850.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Suzhou (China). Association for Computational Linguistics.

Junyou Li, Qi Zhang, and Philip S Yu. 2024. More agents is all you need. *arXiv preprint arXiv:2402.05120*.

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.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

George Michalopoulos, Jean-Philippe Corbeil, Cari Bader, Nate Bodenstab, and Asma Ben Abacha. 2026. Overview of the mediqa-synur 2026 shared task on observation extraction from nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

Shazia Mitha, Jessica Schwartz, Mollie Hobensack, Kenrick Cato, Kyungmi Woo, Arlene Smaldone, and Maxim Topaz. 2023. Natural language processing of nursing notes: an integrative review. *CIN: Computers, Informatics, Nursing*, 41(6):377–384.

AJ Moy, J Adler-Milstein, et al. 2023. Measurement of clinical documentation burden among physicians and nurses using electronic health record audit logs. *Journal of the American Medical Informatics Association*, 30(4):643–652.

OpenAI. 2024. Gpt-4o system card.

S Panchal and P Thakur. 2024. Harnessing the power of natural language processing in nursing services. *International Journal of Advanced Research*, 12(5):154–156.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang,

Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. 2025. Openai gpt-5 system card.

Qwen Team. 2025. Qwen3 technical report.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency

improves chain of thought reasoning in language models. In *ICLR 2023*.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

A. Prompts

To ensure reproducibility, we summarize the prompt templates used in each ACE stage. Prompts follow a strict role–task–constraint–output structure.

A.1. Extraction prompt (core)

The system prompt for both RAG and zero-shot extractors enforces strict JSON schema adherence.

> **ROLE**: You are a clinical documentation expert extracting structured observations from nurse dictations.
>
> **TASK**: Extract all clinical observations mentioned in the transcript and map them to the given ontology.
>
> **OUTPUT FORMAT**: Return a valid JSON object with an `observations` array.
>
> **RULES**:
>
> 1. Extract only observations explicitly mentioned in the transcript.
> 2. Match observation names and IDs to the ontology exactly.
> 3. For `SINGLE_SELECT`, the value must be an exact member of `value_enum`.
> 4. Do not hallucinate observations or values not supported by the transcript.
>
> **ONTOLOGY**: `{schema_str}`

A.2. Adjudication prompt (consensus)

The adjudicator acts as a verifier that resolves disagreements between two candidate drafts.

> **TRANSCRIPT**: `{transcript}`
>
> Below are two candidate extractions from high-performing models.
>
> **DRAFT A** (Model: `{expert_a_name}`, Est. F1: `{expert_a_f1}`): `{draft_a_json}`
>
> **DRAFT B** (Model: `{expert_b_name}`, Est. F1: `{expert_b_f1}`): `{draft_b_json}`
>
> **YOUR TASK**:
>
> 1. Synthesize a single high-fidelity extraction.
> 2. If drafts agree on ID but differ in value, use the transcript to resolve.
> 3. If an observation appears in only one draft, keep it only when explicitly supported by the transcript.
> 4. Ensure all IDs map to valid schema IDs.
>
> Output only the final JSON object.

A.3. Targeted repair prompt

Instead of re-sending the full ontology, the repair prompt provides localized schema-validation feedback to minimize token usage.

> **TRANSCRIPT**: `{transcript}`
>
> **PREVIOUS (INVALID) OBSERVATIONS**: `{invalid_subset_json}`
>
> **ERRORS FOUND**: Observation ID `{id1}`: Value `{val1}` does not match type `SINGLE_SELECT`.
> Observation ID `{id2}`: ID not found in ontology.
>
> Please correct these observations based on transcript evidence and the ontology, and return the full list of valid observations.

B. Additional qualitative cases

We include short qualitative cases to illustrate (i) hallucination removal by adjudication and (ii) value canonicalization by targeted repair. Full transcripts are omitted for space and anonymization.

B.1. Adjudication removes clinically plausible hallucinations

Case A.1 (document ID 0). The dictation describes seizure history and fall-risk awareness, but it does not state several specific structured items. The union stage proposes multiple clinically plausible observations that are unsupported by the transcript; the adjudicator correctly removes them:

- **History of falls** (concept `78`): `Yes` (unsupported),
- **Ambulatory aid** (concept `21`): `walker` (unsupported),

- **Mobility** (concept 89): `Mildly impaired` (unsupported),
- **Cognitive status** (concept 130): `Disoriented to time` (unsupported),
- **Fall risk identification** (concept 107): `High` (unsupported).

This example illustrates why a verifier-style stage is valuable: these items are plausible extrapolations (fall history, mobility limitations, disorientation) but would be unsafe to record as structured observations without explicit evidence in the dictation.

B.2. Targeted repair fixes low-entropy schema/value mismatches

Case A.2 (document ID 5; Use of accessory muscles, concept 38). Before repair, the system outputs `True` for a `SINGLE_SELECT` concept whose allowed values are `Yes/No`. Targeted repair canonicalizes `True`→`Yes`, matching gold.

Case A.3 (document ID 14; Nutrition status, concept 25). Repair projects a near-miss enum phrase "borderline inadequate"→`inadequate`, matching gold.

Case A.4 (document ID 20; Breathing pattern, concept 13). Repair normalizes container type for a `MULTI_SELECT` field: `labored`→`['labored']`, matching gold.

These cases demonstrate that repair is most effective for constrained transformations (boolean/enum canonicalization and list-vs-scalar normalization) rather than re-extraction.

Lakefront AI Ramblers at MEDIQA-SYNUR 2026: Hybrid Retrieval and LLM Verification for Open-Source Schema-Guided Clinical Information Extraction

Michael T. Saban[1,2], Arsalan Yaghoubi[1], Behnaz Eslami[1,2], Samie Tootooni[2], Dmitriy Dligach[1]

[1]Department of Computer Science, Loyola University Chicago; [2]Department of Health Informatics and Data Science, Loyola University Chicago
[1]1052 W. Loyola Ave, Chicago, IL 60660; [2]2160 S. First Avenue, Maywood, IL 60153
{msaban, ayaghoubi, beslami, mtootooni, ddligach}@luc.edu

Abstract

Schema-constrained clinical information extraction requires identifying text-supported observations and outputting exact schema identifiers and values. In the MEDIQA-SYNUR 2026 shared task, synthetic nursing dictations were mapped to structured JSON outputs aligned with a 193-concept clinical schema under strict exact-match evaluation.

We extended the baseline pipeline, which consists of transcript segmentation, schema retrieval, and LLM-based extraction, with hybrid schema retrieval, supervised fine-tuning (SFT) of open-source LLMs, and LLM-based verification. Our hybrid retrieval approach combined dense embeddings with sparse BM25 representations using a convex combination strategy, improving schema coverage to 0.994 recall@60 on the development set. We evaluated GPT-4o, GPT-4o-mini, Llama-3-8B-Instruct, and Llama-3.3-70B-Instruct, applying LoRA-based SFT to open-source models.

On the official test set, our best submitted configuration (Llama-3.3-70B-Instruct-SFT with union voting and GPT-4o-mini verification) achieved 0.711 F1. Post-competition experiments showed that Llama-3-8B-Instruct-SFT (train + dev) reached 0.723 F1 under the same post-processing pipeline. For reference, GPT-4o achieved 0.791 F1 and did not benefit from post-processing. Performance differences across development and test splits further highlight the sensitivity of post-processing strategies to variation across split distribution. Overall, integrating high-recall retrieval, SFT, and LLM verification substantially narrows the performance gap between open- and closed-source models for schema guided clinical extraction.

Keywords: Clinical NLP, Information extraction, Large Language Models (LLMs), Open source, Schema alignment, Hybrid retrieval, Supervised fine-tuning

1. Introduction

Clinical documentation, such as nursing dictations, contains detailed information about patient status, interventions, clinical reasoning, and more (Corbeil et al., 2025; Demsash et al., 2023). However, it is often recorded as unstructured free text. Extracting structured information from these narratives is necessary for tasks such as clinical decision support, patient monitoring, and research (Sezgin et al., 2023).

This work is conducted in the context of the MEDIQA-SYNUR @ ClinicalNLP 2026 shared task, which provides a curated dataset of synthetic nursing dictations paired with gold-standard structured observations corresponding to a predefined extraction schema. Systems are evaluated based on the accuracy of structured JSON outputs at the observation and field levels, using exact and partial matching criteria that require correct concept identifiers, value representations, and correct value types (Corbeil et al., 2025; Michalopoulos et al., 2026). Performance is reported in terms of precision, recall, and F1 score, with systems ranked by overall F1. We follow the official task setup and evaluation protocol without modification.

The original pipeline performs schema filtering using cosine similarity within a retrieval augmented generation framework (Corbeil et al., 2025), essentially relying on dense semantic similarity for selection. Dense representations have become a primary method in retrieval systems (Lin and Lin, 2023). However, prior work has shown that standard dense retrievers can fall behind sparse methods such as BM25 in matching exact lexical cues and rare terminologies, which motivates hybrid approaches that incorporate both semantic and lexical signals (Chen et al., 2022). Hybrid retrieval combining sparse lexical and dense semantic similarity has been studied in information retrieval, where convex combinations of scores consistently outperform either component individually (Bruch et al., 2023; Lin and Lin, 2023; Mandikal and Mooney, 2024). Motivated by this, we apply a convex combination of BM25 and dense embeddings to the schema filtering stage of the MEDIQA-SYNUR pipeline.

Beyond this, we extend the pipeline through supervised fine-tuning of open-source LLMs and evaluation of post-processing strategies, including voting and LLM-based verification. We show that hybrid retrieval substantially improves schema item recall during filtering, while LLM-based verification provides the largest

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 212–221
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

improvements in extraction F1. Together, these components significantly improve structured extraction performance and improve the competitiveness of open-source models in schema-guided clinical information extraction.

2. Methods

We build on the shared task's schema-guided extraction formula and baseline approach by incorporating hybrid schema retrieval and model- and post-processing-level enhancements (**Figure 1**). Each transcript is first segmented into subsets, then processed independently per segment using retrieval-augmented generation (RAG) to select a small subset of schema concepts. A large language model (LLM) produces schema-constrained JSON extractions, which are always followed by deterministic validation. Voting and LLM-based verification are optional experimental post-processing steps that were applied only in the configurations reported for those ablations.

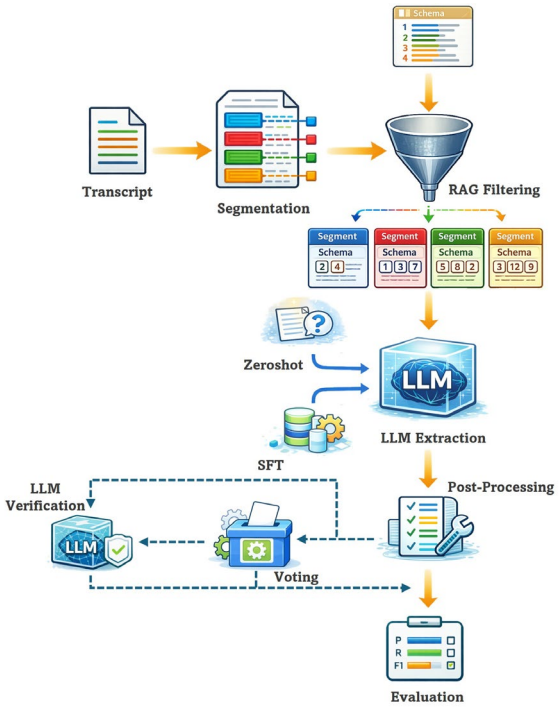


Figure 1. Overview of the schema-guided extraction pipeline. Solid lines indicate the default inference path, while dashed lines denote optional post-processing steps.

2.1 Transcript Segmentation

The MEDIQA-SYNUR transcripts are formatted with repeated "[Clinician]" tags before each span. In this dataset, every span is associated with a single clinician role. We apply a lightweight rule-based segmentation that splits transcripts at each "[Clinician]" tag. Each resulting span is treated as an independent segment for retrieval and extraction.

2.2 Schema-Guided Retrieval

The official schema contains 193 concepts, making it inefficient to include all concepts in-context for every segment. To address this, we use retrieval to select a subset of candidate schema items for each segment.

Dense embeddings are effective for capturing semantic similarity but may miss short lexical signals common in clinical terms (e.g. "Fall risk"). To address this, we adopt a hybrid retrieval strategy that combines dense and sparse representations (Bruch et al., 2023). For a given transcript segment x and schema item s, the final relevance score is calculated as a convex combination of normalized dense and sparse similarity scores:

$$\text{score}(x,s) = \alpha \cdot \hat{S}_{\text{dense}}(x,s) + (1-\alpha) \cdot \hat{S}_{\text{sparse}}(x,s)$$

where $\hat{S}$ denotes min-max normalized similarity scores and $\alpha \in [0,1]$ controls the relative weighting of semantic versus lexical similarity. For each segment, we select the top-N Schema items and provide them to the extraction model as constrained context.

2.3 LLM-Based Extraction

Given a transcript segment and its retrieved schema subset, we prompted an LLM to extract structured clinical observations in a strict JSON format. The prompt consisted of a system instruction defining the model's role as an electronic health record (EHR) flowsheet extraction expert and specifying detailed constraints on schema compliance. These constraints included extracting only explicitly stated or clearly implied observations, matching values exactly to schema-defined enumerations when applicable, returning numeric fields without units, formatting multi-select fields as arrays, and avoiding any generation of unsupported observations. The model was instructed to return a JSON object containing an array of observations with the required fields. This extraction step is model-agnostic and is evaluated with both API-based and open-source instruction-tuned LLMs.

2.4 Supervised Fine-Tuning

To better adapt open-source LLMs to schema-constrained extraction, we applied supervised fine-tuning (SFT) using low-rank adaptation (LoRA) (Hu et al., 2022). A challenge is that gold observations are provided at the transcript level, while segmentation is performed afterward for retrieval and extraction. The dataset does not provide alignment between individual observations and specific transcript segments, making segment-level supervision ambiguous. To avoid ambiguous segment-level supervision, we fine-tuned using transcript-level targets with a fixed retrieval context size (top-N) per transcript. At inference time, the same model is applied per segment.

2.5 Post-processing and Validation

Raw LLM outputs can contain formatting errors and unsupported (hallucinated) observations. We applied the following steps:

2.5.1 Schema Validation

The first stage involved deterministic validation and normalization. Model outputs were parsed and checked against the official schema, removing malformed JSON, discarding observations with invalid schema identifiers, and normalizing categorical values to match the allowed set of schema-defined options. For multi-select fields, single values were converted to lists when necessary, and invalid selections were removed.

2.5.2 Voting

The voting procedure was related to self-consistency decoding, which involves sampling multiple stochastic generations from a single model and aggregating them to reduce variance, although we applied it to schema-constrained information extraction rather than chain-of-thought reasoning (Wang et al., 2022).

Specifically, we ran extraction multiple times at a non-zero temperature and aggregated the resulting observation sets. We considered both majority voting, which retains observations only appearing in multiple runs, and union voting, which retains any observation produced across at least one run.

2.5.3 LLM Verification

To reduce unsupported or hallucinated extractions, we introduced a verification stage using a separate LLM. For each transcript segment, the verifier was provided with the original transcript and the set of extracted observations. The model was prompted to assess each observation and return a structured JSON mapping from observation ID to a binary decision ("KEEP" or "REMOVE"). Observations marked "REMOVE" were discarded from the final output. This stage operated independently from the extractor model and did not modify observations that were kept.

3. Experimental Setup

This section describes the datasets, configurations, and evaluation protocol used to assess the system.

3.1 Data

We evaluated our approach on the MEDIQA-SYNUR 2026 dataset, which contains synthetic nursing dictations annotated with gold structured observations derived from a 193-concept schema. The official split includes 122 training transcripts, 101 development transcripts, and 199 test transcripts.

3.2 Model Configurations

We evaluated four extractor LLMs for the extraction stage: (1) *GPT-4o*, (2) *GPT-4o-mini (Achiam et al., 2023)*, (3) *Llama-3-8B-Instruct*, and (4) *Llama-3.3-70B-Instruct (Grattafiori et al., 2024).* For the open-source models, we applied SFT using LoRA when specified. Fine-tuning was performed on the training set for development experiments. For final test submissions, fine-tuning was performed on the combined training and development sets to maximize exposure. Verification models were evaluated separately and could differ from the extraction model. All model comparisons were performed on the development set unless otherwise noted.

Supervised fine-tuning was performed using HuggingFace Transformers (Wolf et al., 2020) and PEFT with LoRA (Hu et al., 2022). The 70B models used 4-bit quantization using QLoRA (Dettmers et al., 2023) to fit in GPU memory, while the 8B models used standard LoRA without quantization. We evaluated multiple fine-tuning configurations across model sizes and training splits. Hyperparameter settings for model configurations are summarized in **Table 1**. Training was performed on a single NVIDIA H200 GPU.

Parameter	**8B**	**70B**
Epochs	5	3
LR	2e-5	1e-5
LoRA r	32	16
LoRA α	64	32
Max Len	4096	4096
Quant	bf16	4-bit

Table 1. SFT hyperparameters for 8B and 70B models. Identical configurations were used for train-only and train + dev runs within each model size.

3.3 Schema Retrieval Configurations

We compared five dense embedding models for schema retrieval: (1) *all-MiniLM-L6-v2*, (2) *all-MiniLM-L12-v2 (Wang et al., 2020)*, (3) *all-mpnet-base-v2 (Song et al., 2020)*, (4) *BAAI/bge-base-en-v1.5 (Xiao et al., 2024)*, and (5) OpenAI's *text-embedding-3-small*. Sparse retrieval was implemented using BM25 (Robertson and Zaragoza, 2009). We evaluated hybrid dense-sparse retrieval (Bruch et al., 2023) across alpha values. Based on development set recall, we selected *BAAI/bge-base-en-v1.5* with α=0.6. For each segment, the top N=60 schema items were retrieved and used as context for extraction.

3.4 Post-processing Configurations

All runs involved deterministic schema validation. Voting and verification were evaluated as optional

post-processing steps and were only applied in configurations explicitly labeled with those steps. For voting, we performed three independent extraction runs with a temperature of T=0.3. We evaluated both majority voting (threshold ≥ 2/3) and union voting (threshold ≥ 1/3). For the LLM verification step, we evaluated *GPT-4o-mini* and *Llama-3.3-70B-Instruct* as verifiers. Verification was applied after schema validation and optional voting.

3.5 Evaluation Metrics

We report precision, recall, and F1 using the official MEDIQA-SYNUR evaluation script. A predicted observation was considered correct if both the schema concept ID and the extracted value exactly matched the reference annotation. All development set experiments were evaluated locally, while test set results were obtained from the official leaderboard.

3.6 Infrastructure

Closed-source models were run via API on a local workstation. Open-source inference and fine-tuning were run on an HPC cluster equipped with NVIDIA H200 GPUs under a consistent software environment.

4. Results

We evaluated the system on the MEDIQA-SYNUR 2026 development set and report official leaderboard results on the test set where available. Unless otherwise noted, all ablation studies were conducted on the development set.

4.1 Schema Retrieval Performance

Table 2 reports schema retrieval recall@60 on the development set, measured as the fraction of gold schema items covered by the top-60 retrieved candidates (per segment). Across all dense retrievers, hybrid dense + sparse retrieval improved coverage relative to dense-only retrieval. The BM25 + BGE-base hybrid achieved the highest recall@60 (0.994), and this configuration was used for all subsequent experiments.

Retriever	Recall@60
MiniLM-L6	0.953
MiniLM-L12	0.960
MPNet	0.938
TE3-small	0.962
BGE-base	0.974
BM25	0.900
Hybrid (TE3-small+BM25)	0.985
Hybrid (BGE-base+BM25)	0.994

Table 2. Development set schema retrieval coverage (recall@60) for dense-only, sparse (BM25), and hybrid dense + sparse retrieval. Recall@60 measures the fraction of gold schema concepts covered by the top-60 retrieved candidates. Hybrid retrieval uses α=0.6.

4.2 Extraction Performance

Table 3 shows development set extraction performance under single-run inference. GPT-4o achieved the highest F1 (0.845), followed by GPT-4o-mini (0.803). Among the open-source models, Llama-3.3-70B-Instruct improved with SFT (0.765 vs. 0.742). The smaller Llama-3-8B-Instruct model benefited substantially from SFT (0.614 vs. 0.549).

Model	Precision	Recall	F1
GPT-4o	0.841	0.850	0.845
GPT-4o-mini	0.826	0.782	0.803
L3-70B-SFT	0.743	0.787	0.765
L3-70B	0.702	0.788	0.742
L3-8B-SFT	0.497	0.804	0.614
L3-8B	0.498	0.611	0.549

Table 3. Extraction performance on the development set under single-run inference. L3-70B and L3-8B denote Llama-3.3-70B-Instruct and Llama-3-8B-Instruct, respectively; SFT indicates supervised fine-tuning with LoRA.

4.3 Post-processing Effects

Table 4 summarizes the best-performing post-processing strategies for each model on the development set. Overall, LLM verification provided the most consistent improvements across models, typically improving precision while maintaining recall. Voting alone provided smaller, model-dependent gains, with majority voting yielding modest improvements and union voting increasing recall at a large cost of precision. Notably, verification increased Llama-3.3-70B-Instruct-SFT from 0.765 to 0.812 F1, and Llama-3-8B-Instruct-SFT from 0.614 to 0.776 F1. **Figure 2** visualizes precision-recall tradeoffs across all evaluated configurations. The combination of voting and LLM verification produced results similar to LLM verification by itself.

Model	GPT-4o-mini	L3-70B-SFT	L3-8B-SFT
Baseline	0.803	0.765	0.614
Best Voting	0.813	0.799	0.69
Best Verify	0.817	0.812	0.776
Union + Verify	0.82	0.8	0.723
Majority + Verify	0.818	0.811	0.77

Table 4. Summary of development set extraction performance (F1) for the single-run baseline and the best-performing voting, verification, and combined post-processing configurations. L3-70B and L3-8B denote Llama-3.3-70B-Instruct and Llama-3-8B-Instruct, respectively; SFT indicates supervised fine-tuning with LoRA.

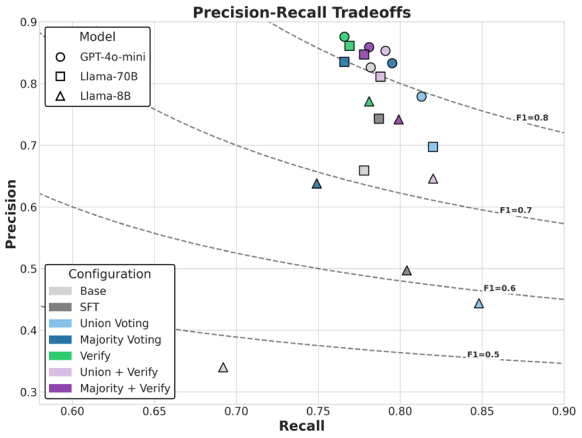


Figure 2. Precision-recall tradeoffs for all models and post-processing configurations on the development set. Shapes indicate model scale; colors indicate configuration; dashed curves show iso-F1 contours.

4.4 Test Set Results

Following the shared task, the organizers released gold labels for the test set, which enabled additional post-hoc evaluation beyond the official leaderboard submissions. **Table 5** reports the performance of our official submissions, all based on Llama-3.3-70B-Instruct-SFT (train + dev) with various post-processing strategies. Relative to the base SFT system (F1 0.680), both voting and verification improved performance, with union + GPT-4o-mini verification achieving the highest F1 among the submitted configurations (0.711).

Table 6 reports additional experiments conducted after the competition using the released test labels. These include base inference with GPT-4o, GPT-4o-mini, and Llama-3-8B-Instruct-SFT (train + dev), along with verification-only and voting + verification variants of the SFT model. These experiments were run once to characterize how post-processing transfers to the test distribution.

Across experiments, voting and verification remained beneficial overall, but relative performance differed across the development and test sets. In particular, the combined voting + verification increased performance more on the test set than verification-only.

Configuration	Precision	Recall	F1
L3-70B-SFT	0.723	0.643	0.680
+ Majority Vote	0.742	0.646	0.691
+ Union Vote	0.648	0.671	0.660
+ Union + GPT-4o-mini Verify	0.768	0.661	0.711
+ Majority + GPT-4o-mini Verify	0.802	0.637	0.710

Table 5. Official test set performance for configurations submitted to the MEDIQA-SYNUR competition. Results are reported for Llama-3.3-70B-Instruct-SFT (train + dev).

Configuration	Precision	Recall	F1
GPT-4o	0.768	0.816	0.791
GPT-4o-mini	0.797	0.586	0.675
L3-8B-SFT	0.733	0.656	0.693
L3-8B-SFT + L3-70B	0.764	0.652	0.704
L3-8B-SFT+ Union + GPT-4o-mini Verify	0.762	0.688	0.723

Table 6. Additional test set performance for configurations run after the MEDIQA-SYNUR competition.

5. Discussion

This study demonstrates that performance in schema-constrained clinical extraction is not exclusively dictated by raw language modeling capacity, but is also influenced by retrieval coverage, model decision supervision, and post-processing. Across our experiments, we found that schema selection defines the feasible context, fine-tuning improves overall output, and verification provides a more stable method of correcting LLM output than stochastic voting.

5.1 Task Difficulty in Schema-Constrained Clinical Extraction

The MEDIQA-SYNUR task highlights the difficulty of schema-constrained extraction because systems must do more than identify clinically relevant content in free text. They must also normalize each extracted observation to a schema by selecting the exact concept identifier and outputting a valid value representation (Fu et al., 2020). Under exact match scoring, errors in unsupported observations, wrong concept IDs, or invalid formatting are all penalized, which makes the evaluation very sensitive to hallucinations and variance in output. This is consistent with how exact-match criteria are usually harsher than overlap-based matching in clinical extraction evaluations (Hein et al., 2025).

The precision-recall tradeoffs across models and post-processing strategies (**Figure 2**) reflect this issue, where improvements in recall come at the cost of including unsupported observations, while aggressively filtering predicted observations increases precision but risks discarding valid clinical observations.

5.2 Effectiveness of Hybrid Schema Retrieval

Schema retrieval is critical in this task, as it determines the context available to the extractor LLM. The success of this step represents an upper bound on extraction performance, as concepts not retrieved in this step cannot be predicted, regardless of model capacity. While hybrid retrieval achieved near perfect recall@60, proving the benefit of combined dense and sparse representations (Bruch et al., 2023; Chen et al., 2022; Gao et al., 2021), extraction F1 was still low. This gap suggests that high schema coverage alone is insufficient, and that retrieval must balance coverage and precision to avoid overwhelming the LLM with irrelevant concepts.

5.3 Model Scaling and SFT Effects

Model scale is generally associated with improved performance in language modeling (Kaplan et al., 2020) and closed-source models such as *GPT-4o* provide a practical upper bound for this pipeline.

However, our findings indicate that scale is not the only determinant of performance under schema-constrained extraction. While *Llama-3.3-70B-Instruct* showed good out-of-box performance and further benefitted from SFT, the 8B model had very competitive results after fine-tuning. Interestingly, when both models were fine-tuned on the combined train + dev set, the 8B model achieved comparable or slightly higher test F1 than the 70B model. While this could suggest that effective fine-tuning can compensate for parameter count differences, we did not conduct hyperparameter optimization across model sizes. This would be supported by further evaluation of model scale under matched hyperparameter searches to isolate the effect of model size.

5.4 Verification vs Voting

Across development set experiments, LLM-based verification was the most consistent post-processing strategy, typically improving precision while maintaining recall. Voting resulted in more variable effects, where majority voting modestly improved performance for some models, but union voting traded precision for minor improvements in recall.

From a practical standpoint, verification is computationally more efficient. The task itself is much smaller compared to the full extraction prompts, so it requires fewer tokens and can be performed using a smaller or cheaper verifier model. Voting requires multiple full extraction passes, making it computationally expensive and less predictable.

Importantly, both strategies improved performance relative to single-run extraction. However, neither approach was completely superior across all settings. The benefit of verification and voting appears to depend on model size, data distribution, and task characteristics. Broader evaluation across datasets and extraction settings Is required to determine how consistently these patterns generalize.

5.5 Open-source competitiveness and deployment relevance

A key finding in this work is that open-source systems can approach or exceed the performance of some closed-source models when combining fine-tuning and verification. On the development set, *Llama-3.3-70B-Instruct* with SFT and verification surpassed the *GPT-4o-mini* baseline and approached *GPT-4o* performance. These results indicate that while proprietary models such as GPT-4o have strong performance without additional supervision, the performance gap between open and closed models narrows noticeably when retrieval, fine-tuning, and verification are optimized for open-source models. This is important when considering the advantages open-source provides in healthcare settings.

Open-source models allow for local deployment, better transparency, and reproducibility, which are all crucial in healthcare settings that are guided by regulatory and infrastructure constraints (Bernal and Mazo, 2022; Riedemann et al., 2024). Open-source alternatives are increasingly viable for schema-guided clinical extraction when combined with the appropriate retrieval and post-processing strategies.

5.6 Generalization Gap on the Test Set

Despite the strong development set performance, test set results had a roughly 0.1 F1 drop, which was primarily due to reduced recall, as precision remained relatively high. Several factors may contribute to this gap. First, differences in annotators across data may create distributional differences across data splits in aspects such as dictation style, concept frequency, or schema-value combinations. Second, the post-processing strategies in this pipeline inherently favored precision, so there were limited opportunities to recover missed observations. Third, the supervision mismatch introduced by transcript-level fine-tuning may have disproportionately affected segment-level evaluation. Gold labels were aligned at the transcript-level, while inference operates per-segment, which increased the likelihood that relevant schema items were excluded from a transcript's top-60 retrieval context even if they appeared in the gold labels.

The generalization gap was not limited to the overall decrease in F1, as the effectiveness of post-processing strategies also shifted. The ranking of verifier models shifted between development and test, and the magnitude of performance from voting + verification differed across these splits. Additionally, while the 70B model clearly outperformed the 8B model when trained on train-only data and evaluated on development, this reversed when both were trained on train + dev data and evaluated on test. Given these results, we can assume that these post-processing strategies and model scaling effects are sensitive to distributional differences and that configuration choices validated on development data may not transfer to unseen data. Post-processing should be viewed as generally beneficial but not universally optimal.

5.7 Future Work

Future work could focus on reducing retrieval noise through adaptive retrieval by dynamically truncating the candidate schema set per segment using score-based cutoffs. Given the high recall@60 achieved by hybrid retrieval, truncation could preserve coverage while providing a more exact context for extraction and reducing the chance of hallucinations. Segment-aligned supervision via heuristics or generated 'silver' segment labels could improve fine-tuning. Verification could be extended with calibrated confidence thresholds, ensemble verifiers, or "unknown" options rather than hard rejection. A broader exploration of newer instruction-tuned models and domain-specific verifiers could close the gap between open- and closed-source systems. Finally, external evaluation on real nurse dictations or other real-world clinical speech transcripts to assess generalizability beyond the synthetic shared-task setting is an important next step.

6. Conclusion

In this work, we investigated schema-constrained information extraction in the MEDIQA-SYNUR 2026 shared task, analyzing how hybrid schema retrieval, supervised fine-tuning, and post processing strategies interact under strict exact-match evaluation. Hybrid dense + sparse retrieval was a critical prerequisite, achieving near-perfect schema coverage, but it was not sufficient on its own. Downstream modeling choices and validation mechanisms play equally important roles in determining final extraction quality.

On the development set, LLM-based verification was the most reliable post-processing strategy, consistently improving precision while preserving overall recall. However, test-set analyses with released labels showed that the contributions of post-processing strategies can change across splits, suggesting sensitivity to distributional differences and post-processing configuration.

Overall, these results demonstrate that effective schema-constrained clinical extraction depends on tightly integrated retrieval, fine-tuning, and post-processing rather than model scale alone. Future work should focus on reducing retrieval noise, improving segment-aligned supervision, conducting systematic hyperparameter selection across model sizes, and calibrated verification strategies to improve system robustness under distribution shift and across related extraction tasks.

7. Limitations

There are several limitations worth noting. First, the dataset consists of synthetic nursing dictations provided as part of the shared task, so we were not able to evaluate on real-world nurse dictations. As a result, the generalizability of these findings to real clinical speech and documentation workflows is uncertain, as synthetic data may not completely capture the variability and ambiguity present in real-world clinical documentation. Although the schema and annotations were carefully constructed, inconsistencies in value formatting and concepts can still affect exact-match evaluation. Second, supervised fine-tuning relied on transcript-level targets due to the lack of segment-aligned annotations, which may limit the effectiveness of segment-level inference. Third, we did not perform any systematic hyperparameter optimization. These results should be interpreted as exploratory rather than as a fully tuned upper bound. Additional limitations include the fixed retrieval context size ($N=60$) and limited voting depth (three runs).

8. Acknowledgments

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM012973 and by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number 5R01HL173037-02. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

9. Bibliographical References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & Anadkat, S. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bernal, J., & Mazo, C. (2022). Transparency of artificial intelligence in healthcare: insights from

professionals in computing and healthcare worldwide. *Applied Sciences*, *12*(20), 10228.

Bruch, S., Gai, S., & Ingber, A. (2023). An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems*, *42*(1), 1-35.

Chen, X., Lakhotia, K., Oguz, B., Gupta, A., Lewis, P., Peshterliev, S., Mehdad, Y., Gupta, S., & Yih, W.-t. (2022). Salient phrase aware dense retrieval: can a dense retriever imitate a sparse one? In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 250-262.

Corbeil, J.-P., Abacha, A. B., Michalopoulos, G., Swazinna, P., Del-Agua, M., Tremblay, J., Daniel, A. J., Bader, C., Cho, K., & Krishnan, P. (2025). Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 859-870.

Demsash, A. W., Kassie, S. Y., Dubale, A. T., Chereka, A. A., Ngusie, H. S., Hunde, M. K., Emanu, M. D., Shibabaw, A. A., & Walle, A. D. (2023). Health professionals' routine practice documentation and its associated factors in a resource-limited setting: a cross-sectional study. *BMJ health & care informatics*, *30*(1), e100699.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, *36*, 10088-10115.

Fu, S., Chen, D., He, H., Liu, S., Moon, S., Peterson, K. J., Shen, F., Wang, L., Wang, Y., & Wen, A. (2020). Clinical concept extraction: a methodology review. *Journal of biomedical informatics*, *109*, 103526.

Gao, L., Dai, Z., & Callan, J. (2021). COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. *arXiv preprint arXiv:2104.07186*.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., & Vaughan, A. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hein, D., Christie, A., Holcomb, M., Xie, B., Jain, A., Vento, J., Rakheja, N., Shakur, A. H., Christley, S., & Cowell, L. G. (2025). Prompts to table: specification and iterative refinement for clinical information extraction with large language models. *medRxiv*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations (ICLR), pages.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Lin, S.-C., & Lin, J. (2023). A dense representation framework for lexical and semantic matching. *ACM Transactions on Information Systems*, *41*(4), 1-29.

Mandikal, P., & Mooney, R. (2024). Sparse meets dense: A hybrid approach to enhance scientific document retrieval. *arXiv preprint arXiv:2401.04055*.

Michalopoulos, G., Corbeil, J.-P., Bader, C., Bodenstab, N., & Ben Abacha, A. (2026). Overview of the MEDIQA-SYNUR 2026 Shared Task on Observation Extraction from Nurse Dictations. In Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, pages, Palma, Mallorca, Spain.

Riedemann, L., Labonne, M., & Gilbert, S. (2024). The path forward for large language models in medicine is open. *npj Digital Medicine*, *7*(1), 339.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, *3*(4), 333-389.

Sezgin, E., Hussain, S.-A., Rust, S., & Huang, Y. (2023). Extracting medical information from free-text and unstructured patient-generated health data using natural language processing methods: feasibility study with real-world data. *JMIR Formative Research*, *7*, e43014.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, *33*, 16857-16867.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, *33*, 5776-5788.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., & Funtowicz, M. (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.

Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., & Nie, J.-Y. (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, pages 641-649.

Appendix A: Detailed Experimental Results

This appendix provides the full experimental results summarized in Table 4 of the main text. All results are reported on the development set unless otherwise noted.

A.1 Voting Strategies

Model	P	R	F1
GPT-4o-mini-M	0.836	0.786	0.810
GPT-4o-mini-U	0.824	0.796	0.810
L3-70B-SFT-M	0.750	0.791	0.770
L3-70B-SFT-U	0.648	0.805	0.718
L3-8B-SFT-M	0.546	0.816	0.654
L3-8B-SFT-U	0.449	0.837	0.585

Table A1. Effect of voting strategies on extraction performance (development set).

A.2 LLM Verification

Model	Verifier	P	R	F1
GPT-4o	GPT-4o	0.865	0.823	0.844
GPT-4o	GPT-4o-mini	0.865	0.827	0.846
GPT-4o	L3-70B	0.863	0.827	0.845
GPT-4o	L3-8B	0.859	0.825	0.841
GPT-4o-mini	GPT-4o-mini	0.870	0.763	0.813
GPT-4o-mini	L3-70B	0.876	0.766	0.817
GPT-4o-mini	L3-8B	0.868	0.756	0.808
L3-70B-SFT	GPT-4o-mini	0.835	0.766	0.799
L3-70B-SFT	L3-70B	0.861	0.769	0.812
L3-70B-SFT	L3-8B	0.854	0.752	0.800
L3-8B-SFT	GPT-4o-mini	0.698	0.786	0.739
L3-8B-SFT	L3-70B	0.771	0.781	0.776
L3-8B-SFT	L3-8B	0.712	0.762	0.736

Table A2. Extraction performance with and without LLM verification on the development set.

A.3 Combined Voting + Verification

Model	P	R	F1
GPT-4o-mini-U	0.853	0.791	0.820
GPT-4o-mini-M	0.859	0.781	0.818
L3-70B-SFT-U	0.811	0.788	0.800
L3-70B-SFT-M	0.847	0.778	0.811
L3-8B-SFT-U	0.646	0.820	0.723
L3-8B-SFT-M	0.742	0.799	0.770

Table A3. Pipeline performance for selected configurations on the development set. Llama-3.3-70B-Instruct was used as the verifier in all configurations.

Appendix B: LLM Prompts

This appendix provides the complete prompts used for LLM-based extraction and verification, as described in sections 2.3 and 2.5.2 of the main text. Template variables (scehma, transcript, observations) are replaced at runtime with the corresponding inputs.

B.1 Extraction Prompts

System Prompt

```
You are an expert at medical electronic health record (EHR) flowsheet analysis.

Your task is to extract clinical observations from nurse dictation transcripts and structure them according to a provided schema.

Guidelines:
1. Extract ONLY observations that are explicitly mentioned or clearly implied in the transcript
2. Match extracted values to the schema's value enum when applicable (use exact matches)
3. For NUMERIC types, extract only the numeric value (no units)
4. For MULTI SELECT types, return an array of applicable values
5. For STRING types, extract the relevant text as mentioned
6. For SINGLE SELECT types, choose the single best matching option from value enum
7. Do NOT fabricate or infer observations that are not in the transcript
8. If unsure, do NOT include the observation

Output Format:
Return a JSON object with key "observations" containing an array. Each observation must have:
- "id": The schema concept ID (string)
- "name": The schema concept name (string)
- "value type": The type from schema (string)
- "value": The extracted value (type depends on value type)

For MULTI SELECT, value should be an array of strings.
For NUMERIC, value should be a number.
For STRING and SINGLE SELECT, value should be a string.
```

User Prompt

```
SCHEMA:
{schema}

TRANSCRIPT:
{transcript}

Extract all clinical observations from the
TRANSCRIPT that match concepts in the SCHEMA.
Return a JSON object with key "observations"
containing an array of extracted
observations.

OUTPUT:
```

B.2 Verification Prompts

System Prompt

```
You are a clinical documentation
verification expert.

Your task is to verify whether extracted
clinical observations are actually supported
by the source nursing transcript.

For each observation, determine:
- KEEP: The observation is directly stated
or clearly implied in the transcript
- REMOVE: The observation is NOT supported,
was incorrectly inferred, or fabricated

Be conservative - only mark REMOVE if you are
confident the observation is not supported.
Small variations in wording are acceptable
if the meaning matches.
```

User Prompt

```
TRANSCRIPT:
{transcript}

EXTRACTED OBSERVATIONS TO VERIFY:
{observations}

For each observation ID, respond with either
"KEEP" or "REMOVE".
Return a JSON object mapping observation IDs
to decisions.

Example response format:
{"40": "KEEP", "56": "REMOVE", "89": "KEEP"}

Your verification:
```

JMedWiC: A Japanese Word-in-Context Dataset in the Medical Domain

Koki Horiguchi[1] **Seiji Sugiyama**[1] **Tomoyuki Kajiwara**[1,2]
Shoko Wakamiya[3] **Eiji Aramaki**[3]
Ehime University[1] The University of Osaka[2] Nara Institute of Science and Technology[3]
{horiguchi@ai., sugiyama@ai., kajiwara@}cs.ehime-u.ac.jp {wakamiya, aramaki}@is.naist.jp

Abstract

We release JMedWiC, a Japanese dataset for Word-in-Context (WiC) tasks specifically tailored to the medical domain. To address the challenge of word sense disambiguation, where the meaning of a word varies depending on its context, previous research has developed WiC datasets to evaluate word sense identity by determining whether a target word shares the same sense across two given contexts. In the medical domain, the misinterpretation of word senses can hinder the accurate comprehension of medical information; however, there is currently no Japanese WiC dataset specialized for this domain. Moreover, existing WiC datasets have been constructed using lexical resources with sense inventories, such as WordNet and UMLS, but such resources are not sufficiently developed for Japanese. Therefore, we construct a Japanese WiC dataset in the medical domain by manually annotating sense-identity labels for target words in context pairs automatically extracted from a large-scale corpus, without relying on lexical resources.

Keywords: Word-in-Context, Medical NLP, Japanese Dataset

1. Introduction

Word sense ambiguity, where a word takes on different meanings depending on the context, is one of the causes of performance degradation in downstream tasks such as machine translation (Chan et al., 2007) and information retrieval (Stokoe et al., 2003). To address this issue, research on word sense disambiguation (Navigli, 2009), which assigns appropriate sense labels to words in a context, has been conducted. In recent years, as a framework that does not rely on a sense label inventory, the Word-in-Context (WiC) task (Pilehvar and Camacho-Collados, 2019) has been proposed, which determines whether a target word has the same meaning in two different contexts. WiC datasets have been developed primarily for English (Breit et al., 2021; Pilehvar and Camacho-Collados, 2019; Raganato et al., 2020).

In the medical domain, many words share the same spelling as common words but have different meanings in specialized contexts. For example, the word “pocket” refers to a bag-like part sewn into clothing in everyday conversation. In surgery, it denotes a wound cavity larger than a skin defect, and in dentistry, it refers to the groove formed between a tooth and the surrounding gum. Such polysemy specific to the medical domain is not fully covered by existing WiC datasets. Therefore, while a WiC dataset specific to the medical domain has been constructed for English (Rouhizadeh et al., 2024), no such dataset exists for Japanese or other languages.

Furthermore, many existing WiC datasets are constructed based on lexical resources that explicitly define sense inventories, such as WordNet (Miller, 1994) and UMLS (Bodenreider, 2004), by extracting context pairs in which the target word has the same sense and context pairs in which it does not. In contrast, Japanese sense inventories specific to the medical domain are not sufficiently developed, and it is difficult to construct datasets using a similar approach.

In this study, we leverage contextualized word representations generated by BERT (Devlin et al., 2019) to construct a Japanese WiC dataset for the medical domain. Contextualized word representations are capable of capturing semantic differences of the same word based on its context as continuous representations. Leveraging this property, we automatically extract synonymous pairs and non-synonymous pairs and then perform manual annotation to construct JMedWiC[1], a Japanese WiC dataset for the medical domain. Our JMedWiC consists of two subsets, medical and general, each containing 1,000 context pairs.

2. Related Work

WiC (Pilehvar and Camacho-Collados, 2019) is a task that determines whether a target word has the same meaning in two different contexts, and datasets have been developed primarily for English (Breit et al., 2021; Pilehvar and Camacho-Collados, 2019; Raganato et al., 2020). The WiC datasets are centered on the lexical resource WordNet (Miller, 1994), which organizes words by sense units (synsets), and are constructed based on whether the target words share the same sense ID. In the medical domain, BioWiC (Rouhizadeh et al.,

[1] https://github.com/EhimeNLP/JMedWiC

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 222–227
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

	Target	Contexts (A / B)	Label
medical	痛み	**A:** この頃より肘と右肩の**痛み**に悩まされ始める。 (From around this time, he began to suffer from **pain** in his elbow and right shoulder.) **B:** エナメル質のう蝕では**痛み**を感じることはない。 (**Pain** is not felt in dental caries of the enamel.)	True
	胸壁	**A:** **胸壁**の良性腫瘍として最も頻度の高いのはどれか。 (Which is the most common benign tumor of the **chest wall**?) **B:** 銀地に、赤い 2 本の**胸壁**のある塔のある城。 (On a silver field, a castle with a tower featuring two red **breastworks**.)	False
general	四季	**A:** **四季**を通じての自然鑑賞地ともなっている。 (It is also a place for appreciating nature throughout the **four seasons**.) **B:** 大陸性気候に属する気候は比較的穏やかで、**四季**もある。 (Climates classified as continental are relatively temperate and feature **four seasons**.)	True
	足取り	**A:** PDF の初期の普及の**足取り**は緩やかなものであった。 (The early **adoption** of PDF was gradual.) **B:** 得意手は右四つ、押し、**足取り**。 (His preferred techniques are a right-hand inside grip, pushing, and **leg picks**.)	False

Table 1: Examples from JMedWiC

2024) has been constructed based on the Unified Medical Language System (UMLS) (Bodenreider, 2004), which integrates medical and biomedical concept systems, and is built according to whether the target word shares the same concept ID.

All of these existing datasets rely on lexical resources that explicitly maintain sense inventories, and in domains where such resources are unavailable, constructing datasets using the same methodology is challenging. In the Japanese medical domain, lexical resources such as WordNet or UMLS have not been sufficiently developed, and consequently, no WiC dataset has been constructed.

3. JMedWiC Datasets

To evaluate the ability to determine word sense identity in Japanese, we constructed JMedWiC, a WiC-format dataset. JMedWiC consists of two subsets: `medical`, which targets medical terms, and `general`, which targets general-domain terms. In this study, we select the target vocabulary for evaluation in each subset and collect their contextual occurrences from two different text corpora. We then automatically extract context pairs based on the semantic similarity of the target word and determine the final labels through manual annotation. Table 1 presents examples from our JMedWiC.

3.1. Target Vocabulary and Contexts

For the `medical` subset, we use JMED-DICT mini[2] as a vocabulary list for the medical domain. JMED-DICT is a dictionary that systematically collects vocabulary in the Japanese medical domain and consists of three sub-datasets: BODY, MEDICATION, and DISEASE. From this dictionary, we extracted 9,257 terms based on frequency information, and selected 6,476 terms consisting of 5 characters or fewer as target words. For the `general` subset, we extracted 7,500 content words from the high-frequency entries of the BCCWJ vocabulary list[3], and selected 7,395 words consisting of 5 characters or fewer as target words.

In this study, we use Wikipedia[4] as a corpus covering a wide range of general-domain contexts. In addition, to ensure the inclusion of medical-domain contexts, we collected medical texts from the online medical encyclopedia, the MSD Manuals[5]. For each corpus, we applied a rule-based sentence segmentation method[6]. For the `medical` subset, sentences containing the target words were extracted from both Wikipedia and the MSD Manuals, whereas for the `general` subset, such sentences were extracted from Wikipedia. To accurately identify the occurrence positions of the target words, we applied word segmentation using MeCab (Kudo et al., 2004)[7], incorporating a medical dictionary[8]. In addition, to exclude extremely short or overly long

[2] https://sip3-d2.naist.jp/jmed-dict.html

[3] Short-unit word list ver.1.0 of the Balanced Corpus of Contemporary Written Japanese (BCCWJ).

[4] https://huggingface.co/datasets/range3/wiki40b-ja

[5] https://www.msdmanuals.com

[6] https://github.com/wwwcojp/ja_sentence_segmenter

[7] https://taku910.github.io/mecab/

[8] https://sociocom.naist.jp/j-meddic-for-mecab/

sentences, only sentences with lengths between 10 and 50 characters were retained.

3.2. Context Pairing

To efficiently extract candidate pairs for manual annotation, we propose a context-pairing method based on the cosine similarity of contextualized word embeddings. Since contextualized word embeddings represent the same word differently depending on its context, the semantic closeness of a word can be estimated based on the cosine similarity between its vectors in two different contexts.

In this method, each context is fed into a masked language model, and the hidden-state vector corresponding to the target word token is used as the contextualized word embedding. If the target word is segmented into multiple subwords, the mean of their vectors is used. The cosine similarity between these vectors is computed for context pairs containing the same target word. We extracted pseudo-synonymous pairs and pseudo-non-synonymous pairs using the predetermined thresholds θ_{high} and θ_{low}.

3.3. Automatic Construction of Dataset

To select a masked language model suitable for context pairing, we employed a word alignment task to measure the models' ability to identify semantic correspondences between words. Four Japanese masked language models were evaluated: BERT[9] (Devlin et al., 2019), RoBERTa[10] (Liu et al., 2019), ModernBERT[11] (Warner et al., 2025), and JMedRoBERTa[12]. For evaluation, we used 300 sentence pairs extracted from the Japanese evaluation set of the medical text simplification parallel corpus[13] (Horiguchi et al., 2025), which consists of semantically aligned complex and simplified medical sentences. For each sentence, word segmentation was performed using MeCab (Kudo et al., 2004) augmented with the medical dictionary[7], and the first author manually annotated the word alignments. For the alignment method, we used SimAlign (Jalili Sabet et al., 2020), which pairs word tokens that maximize the cosine similarity of their contextualized embeddings[14], and performance was evaluated using the F-score.

[9] https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

[10] https://huggingface.co/nlp-waseda/roberta-base-japanese

[11] https://huggingface.co/sbintuitions/modernbert-ja-130m

[12] https://huggingface.co/alabnii/jmedroberta-base-sentencepiece

[13] https://github.com/EhimeNLP/MultiMSDcorpus

[14] SimAlign parameters: bpe, argmax, distortions=0.1

	Precision	Recall	F-score
BERT	65.51	76.64	**70.64**
RoBERTa	65.63	72.08	68.70
ModernBERT	64.31	76.71	69.96
JMedRoBERTa	53.30	77.76	63.24

Table 2: Results of Word Alignment Using SimAlign

	True	False	Total
medical	570 (700)	430 (300)	1,000
general	597 (500)	403 (500)	1,000

Table 3: Label distribution of JMedWiC (parentheses show automatic extraction counts).

As shown in Table 2, BERT achieved the highest F-score. Therefore, in this study, we adopt BERT and perform the context pairing described in Section 3.2 on the set of contexts extracted in Section 3.1. To avoid pairing non-medical contexts within the `medical` subset, we imposed the condition that at least one context in each pair must be derived from the MSD Manuals. In context pairing, pairs with a cosine similarity of $0.75 \leq \theta_{\text{high}} \leq 0.80$ were classified as pseudo-synonymous pairs, while those with a cosine similarity less than $\theta_{\text{low}} = 0.6$ were classified as pseudo-non-synonymous pairs. From the candidate set obtained through context pairing, we selected a total of 1,000 pairs for the `medical` subset, consisting of 700 pseudo-synonymous pairs and 300 pseudo-non-synonymous pairs, and a total of 1,000 pairs for the `general` subset, consisting of 500 pseudo-synonymous pairs and 500 pseudo-non-synonymous pairs. The lower proportion of non-synonymous pairs in the `medical` subset is due to the limited size of the MSD Manuals compared to Wikipedia, which restricted the number of contexts available for extraction. Since the contexts collected from the MSD Manuals are subject to redistribution restrictions, the corresponding contexts among the selected 1,000 pairs were rephrased while preserving their meaning using MedQwen-72b[15] (Kawakami et al., 2025), a large language model specialized in the medical domain.

3.4. Synonymy Annotation

For the collected context pairs, one medical professional annotated the `medical` subset, and one student annotated the `general` subset, indicating whether the target word had the same meaning in each pair. As a result of the annotation, WiC-format datasets consisting of 1,000 pairs each were con-

[15] https://huggingface.co/pfnet/Preferred-MedLLM-Qwen-72B

		medical			general		
		Precision	Recall	F-score	Precision	Recall	F-score
Masked Language Model	BERT	0.693	0.868	0.771	0.839	0.760	0.798
	RoBERTa	0.658	0.854	0.744	0.598	0.980	0.743
	ModernBERT	0.628	0.875	0.732	0.598	0.993	0.746
	JMedRoBERTa	0.567	0.956	0.711	0.598	0.997	0.747
Masked Language Model + Clustering	BERT	0.650	0.840	0.733	0.660	0.660	0.660
	RoBERTa	0.572	0.937	0.710	0.574	0.895	0.699
	ModernBERT	0.570	0.984	0.722	0.569	0.993	0.723
	JMedRoBERTa	0.530	0.751	0.621	0.569	0.970	0.717
Large Language Model	Swallow-8b	0.838	0.570	0.678	0.860	0.216	0.345
	llmjp-13b	0.853	0.409	0.553	0.886	0.131	0.228
	Qwen-72b	0.885	0.637	0.741	0.929	0.305	0.459
	MedQwen-72b	0.882	0.563	0.687	0.889	0.362	0.514

Table 4: Evaluation results of word sense identity determination in `medical` and `general` domains.

structed for both subsets (Table 3). The consistency between automatically assigned labels, determined using the cosine similarity thresholds, and the manual annotations was 81.2% for the `medical` subset and 82.3% for the `general` subset.

4. Experiments

We evaluate the performance of three unsupervised methods on JMedWiC.

4.1. Experimental Setup

Masked Language Model Each context is input to BERT, and the subword embeddings corresponding to the target word are averaged to obtain a contextualized word embedding. The cosine similarity between the word embeddings obtained from the two contexts is computed. If the similarity exceeds the threshold θ, the meanings are considered identical; if it is below θ, the meanings are considered different. We used four Japanese BERT models: BERT[9], RoBERTa[10], ModernBERT[11], and JMedRoBERTa[12]. The threshold θ was selected from $\{0.50, 0.55, \ldots, 0.95\}$ as the value that maximized the F-score. It should be noted that BERT was used during dataset construction (Section 3.3), which gives it a favorable condition compared with the other models.

Masked Language Model + Clustering We employ DBSCAN (Ester et al., 1996), a density-based clustering method that does not require a predetermined number of clusters, to evaluate word sense identity. All contexts containing the target words were collected from Wikipedia[4] and the MSD Manuals[5] for the `medical` subset and from Wikipedia for the `general` subset, and contextualized word embeddings were obtained for these contexts. DBSCAN is then applied to the set of contextualized embeddings, including the contexts to be evaluated, and two contexts are considered to have the same meaning if they belong to the same cluster, and different meanings if they belong to different clusters. The same four Japanese masked language models as in the previous section were used.

Large Language Model In zero-shot in-context learning with a large language model (LLM), the model is provided with an instruction, the target word, and a context pair, and is prompted to output whether the target word has the "same" or "different" meaning in the two contexts. The models used include the general-purpose LLMs Swallow-8b[16] (Fujii et al., 2024), llmjp-13b[17] (LLM-jp, 2024), Qwen-72b[18] (Qwen Team, 2025), and the medical-domain LLM MedQwen-72b[15] (Kawakami et al., 2025). The prompt given to the LLMs is as follows:

Prompt

以下の 2 つの文において対象単語の意味が同じかどうかを判断してください。同じ意味であれば「同じ」、異なる意味であれば「違う」と答えてください。(Please determine whether the target word has the same meaning in the following two sentences. If the meaning is the same, answer "same"; if the meaning is different, answer "different.")

4.2. Results

The experimental results are presented in Table 4. BERT exhibited markedly high performance in

[16] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5

[17] https://huggingface.co/llm-jp/llm-jp-3.1-13b-instruct4

[18] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct

both domains, as the cosine similarity was computed based on the same contextualized embeddings used during dataset construction. Excluding BERT, RoBERTa with a cosine similarity threshold achieved the highest performance on the `medical` evaluation, while those of JMedRoBERTa achieved the best performance on the `general` evaluation. Overall, masked language model + clustering demonstrated comparable performance compared with masked language models using optimal cosine similarity thresholds, whereas zero-shot in-context learning with LLM demonstrated a decrease in performance.

When comparing across domains, performance in the `medical` subset tended to surpass that in the `general` subset. Medical terms are often used in specialized contexts, and their range of meanings is relatively limited, which may have made sense disambiguation easier. In contrast, general terms often appear in contexts that are polysemous or figurative, making sense boundaries more ambiguous and thereby increasing the difficulty of sense disambiguation.

Comparing general purpose and medical specialized models on the `medical` subset, no clear advantage was observed for the medical-specialized models; both JMedRoBERTa and MedQwen-72b performed worse than the general-purpose models. This result suggests that what is important for the WiC task may not be domain-specific knowledge about the target words themselves, but rather general linguistic knowledge for identifying sense equivalence from context.

5. Conclusion

In this study, we constructed JMedWiC, consisting of 1,000 pairs each for the medical and general domains, to evaluate the ability to determine context-dependent word sense identity in Japanese medical texts, using contextualized word embeddings for context pairing and manual annotation. Experiments showed that masked language models based on cosine similarity thresholds performed best in both the medical and general domains, outperforming both masked language model + clustering and large language models. In addition, it was revealed that medical-specialized models do not necessarily outperform general-purpose models, and that the task difficulty differs across domains.

Acknowledgements

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on “Integrated Health Care System” Grant Number JPJ012425.

Bibliographical References

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. WiC-TSV: An Evaluation Benchmark for Target Sense Verification of Words in Context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1635–1645.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40.

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*, pages 4171–4186.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page 226 – 231.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In *Proceedings of the First Conference on Language Modeling*.

Koki Horiguchi, Tomoyuki Kajiwara, Takashi Ninomiya, Shoko Wakamiya, and Eiji Aramaki. 2025. MultiMSD: A Corpus for Multilingual Medical Text Simplification from Online Medical References. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9248–9258.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643.

Wataru Kawakami, Keita Suzuki, and Junichiro Iwasawa. 2025. Stabilizing Reasoning in Medical LLMs with Continued Pretraining and Reasoning Preference Optimization. *arXiv:2504.18080*.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Joshi. Mandar, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

LLM-jp. 2024. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. *arXiv:2407.03963*.

George A. Miller. 1994. WordNet: A Lexical Database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1–10:69.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. 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 1267–1273.

Qwen Team. 2025. Qwen2.5 Technical Report. *arXiv:2412.15115*.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7193–7206.

Hossein Rouhizadeh, Irina Nikishina, Anthony Yazdani, Alban Bornet, Boya Zhang, Julien Ehrsam, Christophe Gaudet-Blavignac, Nona Naderi, and Douglas Teodoro. 2024. A Dataset for Evaluating Contextualized Representation of Biomedical Concepts in Language Models. *Scientific Data*, 11(1):455.

Christopher Stokoe, Michael P. Oakes, and John Tait. 2003. Word sense disambiguation in information retrieval revisited. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, page 159 – 166.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

LTRC-Medicom at MEDIQA-SYNUR 2026: Schema-Guided Clinical Information Extraction with Hybrid Clustering-SFT-Verification

Pasumarthy Deepak, Sushvin Marimuthu, Parameswari Krishnamurthy
LTRC, International Institute of Information Technology, Hyderabad, India
pasumarthy.deepak@students.iiit.ac.in,sushvinmarimuthu@gmail.com,param.krishna@iiit.ac.in

Abstract

Extracting structured clinical data from unstructured patient transcripts is challenging due to large target schemas and inherent linguistic ambiguity. We address the extraction of 193 heterogeneous clinical attributes from nursing notes and clinician–patient dialogues, and demonstrate that zero-shot large language models (LLMs) are ineffective in this setting, achieving an F1 score below 0.15 due to context window saturation and hallucination. We propose a four-stage framework that combines semantic schema clustering, role-based chain-of-thought prompting, supervised fine-tuning of Llama-3.1-8B, and transcript-verified post-processing. Our approach achieves an F1 score of 0.66, representing a 4.4x improvement over the baseline, by balancing high recall from generative models with high precision from verification. These results highlight the effectiveness of hybrid pipelines for high-stakes clinical information extraction.

Keywords: clinical information extraction,large language models, schema clustering,supervised fine-tuning,healthcare NLP

1. Introduction

The increasing digitization of healthcare has intensified the demand for efficient clinical documentation, particularly in nursing workflows. Critical patient information is often conveyed during nurse–patient conversations, yet much of it is not systematically captured in structured Electronic Health Records (EHRs).

In mediqa-synur shared task at the Clinical NLP 2026 (George Michalopoulos, 2026), the goal is to advance automated methods for capturing clinically salient information from nursing dialogue data. By improving structured extraction from nurse–patient conversations, the task seeks to reduce nursing documentation burden and enhance the accuracy and completeness of patient records.

Unlike standard Named Entity Recognition (NER) (Xu et al., 2010). which targets generic entities (e.g., “Drug”, “Disease”), this task requires mapping text to a strict schema of 193 constrained fields. Each field enforces structured outputs such as predefined enumerations for “Mobility” (e.g., “Ambulates with assistance” vs. “Bedbound”) or bounded integers for “Pain Score” that trigger downstream protocols. By enabling accurate schema-level extraction from nurse–patient conversations, the task aims to reduce documentation burden and improve the completeness and reliability of patient records.

To address this challenge, we propose a hybrid system that evolves from retrieval-augmented generation (RAG) (Lewis et al., 2020) concepts to supervised learning. By clustering schema fields, we reduce cognitive load and improve task decomposition. Through fine-tuning, we internalize schema constraints within the model. Finally, we incorporate algorithmic verification against the source text to mitigate hallucinations and ensure faithful, schema-compliant extractions.

We systematically explored a progression of modeling strategies with increasing levels of supervision and structural incorporation. A zero-shot baseline performed poorly **(F1 < 15%)**, underscoring the complexity of direct schema-guided extraction. Incorporating schema-aware clustering improved performance to **33% F1**, while systematic prompt engineering further raised it to **45% F1**. Transitioning to supervised fine-tuning (SFT) resulted in a substantial gain, achieving **62% F1**. Finally, augmenting SFT with rule-based post-processing and validation yielded a peak performance of **66% F1**.

2. Related Work

Clinical NER and Information Extraction: Traditional approaches to clinical information extraction have relied heavily on Named Entity Recognition (NER) and relation extraction. Early systems used rule-based methods and classical machine learning (Xu et al., 2010), while modern approaches leverage pretrained transformers such as ClinicalBERT (Alsentzer et al., 2019) and domain-specific models (Yang et al., 2020). However, these methods typically target generic entity types (e.g., medications, diseases) rather than schema-guided extraction with strict constraints.

LLMs for Clinical Tasks: Recent work has explored using large language models for clinical documentation (Agrawal et al., 2022), demonstrating zero-shot and few-shot capabilities. However, these approaches struggle with high-dimensional

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 228–234
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

schemas and exhibit significant hallucination in clinical contexts, motivating our hybrid verification framework.

Retrieval-Augmented Generation: RAG-based methods (Lewis et al., 2020) have shown promise in grounding LLM outputs, but applying RAG to structured extraction from long clinical dialogues remains underexplored. Our semantic clustering can be viewed as a deterministic routing mechanism that shares conceptual similarities with retrieval-based context selection.

3. Dataset

The shared task utilizes SYNUR (Corbeil et al., 2025a), the first dataset specifically designed for structured observation extraction from nursing dictations. The dataset is accompanied by a schema specification in JSON format that defines the complete set of clinical observation fields. Each schema entry includes a unique identifier (id), a field name (name), and a designated value type (value_type). In total, the schema defines 193 distinct clinical variables.

The schema supports four value types: NUMERIC (e.g., vital signs), STRING (free-text notes), SINGLE_SELECT (categorical status variables), and MULTI_SELECT (symptom or condition lists). For categorical fields (SINGLE_SELECT and MULTI_SELECT), the schema further provides a predefined set of allowable values (value_enum), ensuring standardized normalization of extracted observations.

The training and development splits are provided in JSONL format. Each data instance contains a unique identifier (id), the corresponding nursing dictation (transcript), and a list of annotated observations (observations). Each observation entry includes a schema-aligned identifier (id), field name (name), designated value type (value_type), and the normalized field value (value).

The test split is also released in JSONL format and contains only the unique identifier (id) and the associated nursing dictation (transcript), requiring systems to predict the structured observations according to the predefined schema.

4. System Description

Our pipeline consists of four distinct phases, evolved iteratively to address specific failure modes observed in preliminary experiments.

4.1. Phase 1: Semantic Clustering (Hybrid Lexical-LLM Routing)

Motivation: Large Language Models suffer from "Context Overload" when presented with long lists of instructions, often referred to as the "Lost in the Middle" phenomenon (Liu et al., 2023). Our hypothesis was that by decomposing the 193-item schema into semantically cohesive sub-groups, we could maintain the model's attention mechanism on a smaller, relevant set of definitions.

Implementation: Contrary to traditional K-Means approaches which often create incoherent clusters, we implemented a **Two-Tier Hybrid Router** using `Gemma-2-9B-Instruction` (Team, 2024) as a semantic classifier.

- **Tier 1: High-Precision Lexical Routing:** We defined strict prefix rules for known clinical scales to ensure determinism. For example, any item starting with "Glasgow coma score" is automatically routed to the `Neurological` cluster.
- **Tier 2: LLM-Based Semantic Classification:** For items not caught by lexical rules, we employed a strict classification prompt with `Gemma-2-9B`. The model was tasked with assigning each schema item to one of 4 primary buckets (*Numeric, Single Select, Multi Select, String*) and then further into specific clinical domains (e.g., *Cardiovascular, Respiratory, Safety*).

Cluster Definition:

- *Numeric:* Vitals, Cardiac Measures, Fluid Balance.
- *Single Select:* Respiratory, Neurological (GCS), Skin (Braden).
- *Multi Select:* Symptoms, Interventions.
- *String:* Qualitative Observations, Notes.

Inference Architecture: The pipeline processes the transcript in sequential passes. For each pass, a dynamic system prompt is constructed containing only the definitions for that specific domain (e.g., "Neurological Single Selects"). This reduced the effective instruction token count from $\sim$8000 tokens to $<$1000 tokens per pass, preventing context overflow.

Observation: This phase improved F1 from **0.15 to 0.33**, largely by eliminating cross-domain hallucinations (e.g., confusing "Fall Risk" definitions with "Mobility" definitions).

4.2. Phase 2: Systematic Prompt Engineering (The 9-Step Protocol)

Motivation: General-purpose instruction following is insufficient for clinical rigor. Models tend to be "helpful" rather than "accurate", guessing values to complete the task. We required a methodology

to force strict adherence to clinical documentation standards.

Implementation: We designed a 9-Step Chain-of-Thought (CoT) System Prompt based on the persona of a "CDI Specialist", inspired by the winning approach in the MEDIQA-OE 2025 shared task (Corbeil et al., 2025b).

The Full 9-Step Template: shown in Table 1

Key Innovation: For high-confusion items identified during development, we injected field-specific hints to improve disambiguation. For ID 14 (Safety Equipment), originally defined only as "Equipment for safety", we appended: "Hint: Bed rails, Mats, Alarms ONLY. Do not extract walkers/canes." For ID 40 (Orientation), we added: "Hint: Look for 'Alert and Oriented x3' or 'A&O x4'." These targeted hints reduced ambiguity for the 12 most frequently confused fields.

Observation: This improved precision to 0.45 while maintaining recall at 0.45, achieving balanced performance. However, the model remained conservative and often missed rare or implicit clinical cues, such as inferring fall risk from behavioral descriptions. This limitation in capturing nuanced patterns motivated Phase 3.

4.3. Phase 3: Supervised Fine-Tuning (The Recall Engine)

Motivation: While prompting provides task-specific guidance at inference time, supervised fine-tuning embeds schema knowledge directly into model parameters. To capture the full nuance of 193 field definitions—especially rare patterns such as specific skin conditions or IV catheter types—we hypothesized that parameter updates would enable more robust extraction than prompting alone.

Implementation: We employed Supervised Fine-Tuning (SFT) using QLoRA (Quantized Low-Rank Adaptation).

Hyperparameters:

- Base Models: Qwen-3-8B (Team, 2025)
- Quantization: 4-bit (NF4).
- LoRA Rank (r): 128 (High rank chosen to capture complex schema relations).
- LoRA Alpha: 256.
- Target Modules: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj.
- Learning Rate: 2e-4.
- Epochs: 3.

Input Format During Training: Each training example was constructed as a conversational turn with the following structure:

- **User message:** Contains the complete nursing transcript along with the *full schema specification* (all 193 field definitions with their types, enumerations, and descriptions).
- **Assistant message:** Contains the gold-standard JSON output with all annotated observations for that transcript.

Schema Presentation Strategy: During fine-tuning, we presented the complete 193-field schema in every training example, rather than using clustered subsets. This differs from our inference approach (Phase 1), where clustering is employed to reduce context load. The rationale is that fine-tuning allows the model to internalize schema knowledge through parameter updates, eliminating the need for clustering during training. By exposing the model to the full schema repeatedly across 237 examples over 3 epochs (711 total exposures), we enable comprehensive schema memorization.

Prompt Template: We applied the same 9-step Chain-of-Thought prompt (described in Phase 2) during fine-tuning. Each training example began with the complete 9-step system prompt, followed by the transcript and full schema, with the model trained to generate the corresponding JSON output. This ensures alignment between training and inference procedures.

Training-Inference Alignment: A key architectural consideration is the difference between our training and inference pipelines. During training, the model observes complete schema presentations to internalize all 193 fields through parameter updates. During inference (Section 5), we employ Phase 1 clustering to reduce context load, as the model has already internalized schema knowledge through fine-tuning. This hybrid approach combines the benefits of comprehensive training with efficient clustered inference, preventing context overflow while maintaining schema coverage.

Observation (Recall Explosion): The SFT model demonstrated a dramatic behavioral shift. It learned to identify implicit clinical cues that the prompted model missed.

- Prompted Model: Saw "patient trying to get OOB" $\rightarrow$ Extracted Nothing.
- SFT Model: Saw "patient trying to get OOB" $\rightarrow$ Extracted "Fall Risk: Impulsive" and "Mobility: Unsteady".
- Recall: Jumped to 0.75.
- Challenge: The model became "trigger-happy", often inferring conditions where none existed (Precision drop to 0.21).

#	Step Name	Prompt Content
1	Role Attribution	You are an expert Clinical Documentation Improvement (CDI) Specialist with 20 years of experience. You are precise, conservative, and strictly adhere to provided schemas.
2	Transcript Definition	The input text is a 'Transcript' of a clinical encounter. It may contain disfluencies, interruptions, and informal language.
3	Task Definition	Your task is to analyze the Transcript and extract specific clinical variables defined in the Schema.
4	Type Definitions	*[Dynamic Section: Inserts rules for* `SINGLE_SELECT`*,* `MULTI_SELECT`*, or* `NUMERIC` *based on current item]*
5	Output JSON Key Definitions	Return a JSON list of objects with keys: `"id"`, `"name"`, `"value"`, `"value_type"`.
6	Reasoning Guidelines (Negative Constraints)	• ABSOLUTELY NO inference of 'Normal' if not explicitly stated. • If the transcript is silent, SKIP the item. • Silence $\neq$ Normalcy.
7	Evidence Requirement	Before extracting, you must locate the specific mention in the transcript.
8	Overall Guidelines	• Accuracy is paramount. Do not round unless necessary. • Strict Enum Matching: The `value` must strictly match one of the provided options.
9	Eliciting JSON Output	`TRANSCRIPT`: {transcript} `SCHEMA`: {schema_subset} Answer with the JSON list only.

Table 1: The 9-Step Chain-of-Thought System Prompt Protocol used for the "CDI Specialist" persona (Corbeil et al., 2025b).

4.4. Phase 4: Transcript-Verified Post-Processing (The Precision Engine)

Motivation: Generative models are probabilistic and inherently prone to hallucination, particularly when trained to maximize recall. To achieve clinical-grade precision, we must constrain the output with deterministic rules. We treat the fine-tuned model as a candidate generator and employ post-processing as a precision-focused verifier.

Implementation: We developed a robust Python-based verification engine.

- **Algorithm 1: Transcript Verification**
  This algorithm ensures that every extracted string actually exists in the source. This filtered out pure hallucinations (e.g., model inventing "Pain: 5/10" when pain was not mentioned).
- **Algorithm 2: Schema-Based Phantom Filter**
  The model often defaults to "Normal" values (e.g., "Skin: Intact") even when the input is silent, violating the schema's requirement for "Null" values. We enforce strict Schema Adherence by utilizing a `STRICT_DENY_LIST`. If item is in a "Nullable" Schema Category (e.g., Skin, Mood) and Value is in a "Normative Set" (e.g., "Intact", "Normal"), the value is automatically rejected to ensure compliance with the Schema's "No Default" policy.
- **Algorithm 3: Enum Snapshot**
  We enforced strict schema compliance using Levenshtein distance.
  - **Model Output:** "Alert"
  - **Schema Options:** ["Alert and oriented x3", "Confused", "Unresponsive"]
  - **Action:** Match "Alert" to "Alert and oriented x3" (Distance < Threshold).
  - **Result:** Corrects lazy outputs to valid schema enumerations.

Observation: This phase was the most critical for precision, raising it from 0.21 to 0.61 without sacrificing the Recall gained in Phase 3.

5. Inference

Architecture: The inference engine iterates through the semantic clusters defined in Phase

1. For each pass, a dynamic system prompt is constructed containing *only* the definitions for that specific domain (e.g., "Neurological Single Selects"). This reduced the effective instruction token count from $\sim$8000 tokens to $<$1000 tokens per pass, preventing context overflow.

$$P(\text{Output} \mid \text{Transcript}, \text{Schema}) \approx \prod_{i=0}^{k} P(\text{Output}_i \mid \text{Transcript}, \text{Cluster}_i) \quad (1)$$

6. Evaluation

We evaluated the pipeline on a held-out test set of clinical transcripts using standard Information Extraction metrics.

- **Precision**: Measure of correctness (Are extracted items real?).
- **Recall**: Measure of completeness (Did we find all items?).
- **F1 Score**: Harmonic mean of Precision and Recall.

We specifically focused on the **Recall-Precision Trade-off**, analyzing how SFT improved Recall while Verification recovered Precision.

7. Results

7.1. Development Set Results

We first present performance on the development set, which guided our design decisions and hyperparameter selection for each phase.

Phase	Method	F1	Rec	Prec
0	Zero-Shot	0.14	0.12	0.16
1	Clustering	0.31	0.69	0.20
2	Prompting	0.43	0.43	0.43
3	SFT (Qwen-3-8B)	0.60	0.73	0.51
4	+ Verification	**0.64**	**0.70**	**0.59**

Table 2: Performance Evolution by Phase on Development Set

7.2. Test Set Results

Final evaluation on the held-out test set confirms the generalization of our approach.

Phase	Method	F1	Rec	Prec
0	Zero-Shot	$<$0.15	Low	Low
1	Clustering	0.33	0.72	0.21
2	Prompting	0.45	0.45	0.45
3	SFT	0.62	0.75	0.53
4	Post-Proc	**0.66**	**0.72**	**0.61**

Table 3: Performance Evolution by Phase

7.3. Analysis and Discussion

Why Clustering Improved Performance (Phase 1): The clustering phase (F1: 0.14 $\rightarrow$ 0.31) directly addresses the "Lost in the Middle" phenomenon documented by Liu et al. (Liu et al., 2023). By decomposing the 193-field schema into 15 semantic clusters, we reduced the effective context from approximately 8,000 tokens to fewer than 1,000 tokens per pass. This prevents catastrophic forgetting and eliminates cross-domain hallucinations—for example, the model no longer confused "Fall Risk" assessments with "Mobility" status. However, clustering alone achieved high recall (0.69) but poor precision (0.20), as the model still lacked explicit guidance on extraction constraints and when to abstain from inference.

Why Prompting Balanced Precision and Recall (Phase 2): The systematic prompting phase (F1: 0.31 $\rightarrow$ 0.43) introduced explicit constraints through our 9-step Chain-of-Thought protocol. The key innovation was the negative constraint explicitly forbidding inference of "Normal" values when the transcript was silent. This principle—"silence does not equal normalcy"—balanced precision and recall at 0.43 each, correcting the precision deficit from Phase 1. However, the prompted model remained conservative and failed to capture implicit clinical cues. For example, it could not infer "Fall Risk" from behavioral descriptions like "trying to get out of bed" without explicit fall risk language, limiting recall potential.

Why SFT Dramatically Boosted Recall (Phase 3): Fine-tuning (F1: 0.43 $\rightarrow$ 0.62) embedded schema knowledge directly into model parameters through 711 training exposures (237 examples $\times$ 3 epochs). The model internalized implicit clinical reasoning patterns from annotated examples, enabling it to recognize subtle indicators. Recall jumped to 0.73 (+70% relative improvement), as the model learned associations like "trying to get OOB" $\rightarrow$ "Fall Risk: Impulsive" and "appears uncomfortable" $\rightarrow$ potential pain indicators. However, this sensitivity came at a cost: precision increased only marginally to 0.51 (+19% relative), as the model occasionally over-inferred conditions not explicitly stated. This demonstrates the fundamental precision-recall trade-off inherent in generative

extraction systems.

Why Post-Processing Recovered Precision (Phase 4): The verification phase (F1: 0.62 → 0.66) applied deterministic rules to filter false positives while preserving true positives. The three-algorithm pipeline increased precision from 0.51 to 0.59 (+16% relative) with only a minor recall decrease (0.73 → 0.70, -4% relative). This demonstrates the effectiveness of treating fine-tuned models as candidate generators followed by rule-based precision verification. The modest recall loss occurs because some valid implicit inferences lack direct textual evidence (e.g., inferring alertness from conversational engagement), causing them to be filtered by transcript verification. However, the substantial precision gain outweighs this minor recall cost.

Key Trade-offs and Optimal Strategy: Our experiments reveal a systematic precision-recall trade-off at each phase. Clustering sacrifices precision for recall by reducing context overload. Prompting balances both but limits recall due to conservatism. Fine-tuning maximizes recall through pattern learning but increases false positives. Post-processing recovers precision with minimal recall loss through verification. The optimal strategy combines all four phases sequentially, achieving both high recall (0.72) and high precision (0.61) by leveraging the complementary strengths of each approach.

7.4. Qualitative Error Analysis

We analyzed the remaining error modes in the final Phase 4 output:

1. **Semantic Ambiguity**: Confusion between related fields (e.g., "Mobility" vs "Activity Level"). The model typically picks the more common one.

2. **Implicit Negation**: The text says "Patient denies pain", but the model extracts "Pain Severity: 0". While clinically true, if the schema requires "Null" for "No Pain", this counts as a False Positive using strict evaluation.

3. **Complex Temporal Logic**: In long transcripts, the patient status changes (e.g., "Alert" → "Confused"). The model sometimes struggles to "overwrite" the earlier state with the latest one.

8. Conclusion

We have presented a robust framework for high-dimensional clinical data extraction. By combining **Semantic Clustering**, **Chain-of-Thought Prompting**, **Supervised Fine-Tuning**, and **Deterministic Verification**, we overcame the limitations of off-the-shelf LLMs. Our system provides a blueprint for deploying Generative AI in high-stakes healthcare environments: **Trust (the model), but Verify (the transcript).**

9. Code Availability

The complete implementation, including inference and training scripts, schema definitions, training logs, and detailed documentation, is available at: GitHub Repository: ltrc-medicom-mediqa-synur-clinicalnlp-2026

10. Limitations & Ethical Considerations

- **Data Bias**: Our fine-tuning dataset was derived from a specific institution's transcript style. The model may underperform on dialects or documentation styles not represented in the training distribution.
- **Privacy**: All experiments used de-identified datasets. However, deploying Generative AI in healthcare requires strict adherence to HIPAA/GDPR. The "Hallucination" risk, while mitigated, is never zero; thus, this system is designed as a "Human-in-the-Loop" assistive tool, not an autonomous coder.
- **Compute Costs**: While inference is optimized, the SFT process requires GPU resources that may be inaccessible to low-resource settings.

11. References

Monica Agrawal et al. 2022. Large language models are zero-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025a. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of*

the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, Suzhou (China). Association for Computational Linguistics.

Jean-Philippe Corbeil, Asma Ben Abacha, Jérôme Tremblay, Phillip Swazinna, Akila Jeeson Daniel, Miguel Del-Agua, and Francois Beaulieu. 2025b. Overview of the mediqa-oe 2025 shared task on medical order extraction from doctor-patient consultations. *arXiv preprint arXiv:2510.26974*.

Cari Bader Nate Bodenstab Asma Ben Abacha George Michalopoulos, Jean-Philippe Corbeil. 2026. Overview of the mediqa-synur 2026 shared task on observation extraction from nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

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*, pages 9459–9474.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts.

Gemma Team. 2024. Gemma.

Qwen Team. 2025. Qwen3 technical report.

Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. 2010. Medex: A medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24.

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942.

SloCal-Net at MEDIQA-Eval 2026: Investigating the Impact of Reasoning and External Context on Medical Answer Grading

Primoz Kocbek[1,2], **Valentina Carbonari**[3], **Pierangelo Veltri**[4], **Pietro Hiram Guzzi**[3], **Gregor Stiglic**[1,5]

[1] University of Maribor, Faculty of Health Sciences, Žitna ulica 15, 2000 Maribor, Slovenia, {primoz.kocbek, gregor.stiglic}@um.si,

[2] University of Ljubljana, Faculty of Medicine, Vrazov trg 2, 1000 Ljubljana, Slovenia,

[3] Magna Graecia University of Catanzaro, Department of Surgical and Medical Sciences, Viale Europa, 88100 Catanzaro, Italy, {valentina.carbonari, hguzzi}@unicz.it

[4] University of Calabria, DIMES, Via P. Bucci, 87036 Rende, Italy, pierangelo.veltri@unical.it,

[5] Usher Institute, The University of Edinburgh, Edinburgh EH16 4UX, UK

Abstract

Automated evaluation of multimodal medical answers is essential for scalable safety assessment, yet it remains difficult to align automatic scores with expert judgment across languages and image modalities. We describe SloCal-Net's systems for the MEDIQA-EVAL 2026 shared task, framing evaluation as rubric-conditioned multimodal judging: the judge receives the question, image(s), candidate answer, and task-specific criteria, and outputs criterion-level scores and an overall rating. Evidence retrieval was initialized using ChatGPT Deep Research, producing a 25-document clinical corpus used for lightweight retrieval-augmented grounding. On the official leaderboard, our best submission (GPT-5-mini with web search and structured scoring) achieved Pearson correlations of 0.466 on English and 0.260 on Chinese expert ratings. In post-competition experiments with open-source judges, the best English Pearson reached 0.272 with GLM-4.6V and 0.212 with Qwen3-VL-30B-Thinking, while Chinese correlations were lower, highlighting remaining gaps in multilingual calibration and image–text grounding.

Keywords: multimodal evaluation, clinical NLP, LLM-as-a-judge

1. Introduction

The integration of large multimodal language models (LLMs) into clinical workflows has led to a paradigm shift in the generation and assessment of medical responses. However, ensuring reliability, safety, and clinical validity remains a major challenge. Conventional evaluation metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) primarily measure lexical overlap and often fail to reflect medical correctness or appropriateness. Similarly, general-purpose vision–language models such as CLIP (Radford et al., 2021) lack the domain specificity required for nuanced interpretation of medical images. In practice, evaluation systems must perform cross-modal reasoning—aligning written responses with visual clinical evidence such as dermatological images or diagnostic scans—while accounting for the fact that multiple expert answers may be clinically acceptable (Yim et al., 2025).

Historically, automated medical evaluation relied on n-gram–based metrics, which correlate poorly with clinician judgment in open-ended settings (Asma Ben Abacha, 2026). This limitation has motivated a shift toward semantic and reasoning-based evaluation frameworks. More recently, the *LLM-as-a-judge* paradigm has emerged, where strong models are prompted to score candidate outputs against structured clinical rubrics (Zheng et al., 2023). While effective for text-only tasks, purely textual judges struggle in multimodal contexts, where visual evidence must be verified against generated claims (Anonymous, 2025). Shared tasks such as MEDIQA-EVAL therefore require evaluators to detect inconsistencies between patient images and candidate responses (Yim et al., 2025b).

Parallel advances in domain-specific multimodal modeling further shape this landscape. Generalist vision–language encoders often fail to capture fine-grained dermatological distinctions (Yan et al., 2025), motivating specialized models such as DermLIP, trained on the Derm1M dataset of clinically aligned dermatology image–text pairs (Yan et al., 2025). Similarly, MedGemma integrates medical vision encoders optimized for radiology and pathology (Sellergren et al., 2025). These models illustrate the growing importance of domain-specific representations for clinically grounded evaluation.

At the same time, newer reasoning-oriented architectures emphasize inference-time deliberation and structured decision-making. Systems such as the GPT-5 family of models incorporate internal planning mechanisms to improve complex multi-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 235–243
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

modal reasoning (Wang et al., 2025). Retrieval-augmented and agentic workflows further attempt to reduce hallucinations by grounding predictions in external evidence (OpenAI, 2025). Platform-level efforts, including large-scale clinical AI ecosystems and evaluation frameworks such as ChatGPT Health highlight the increasing emphasis on safety, alignment, and real-world deployment constraints.

Within this context, the SloCal-Net team approach adopts a pragmatic LLM-judge strategy for multimodal clinical evaluation. We develop prompt-based judges and study two design dimensions that directly affect performance: (i) per-dimension scoring versus structured per-case evaluation, and (ii) the role of retrieval-augmented grounding. We report official shared-task results and conduct additional analyses with open-weight multimodal judges to assess the remaining performance gap between proprietary and open systems (Asma Ben Abacha, 2026).

2. Task and Data

The MEDIQA-EVAL 2026 shared task addresses this need by providing a benchmark for the automatic evaluation of multimodal medical responses, building on recent work on evaluating medical open-ended question answering such as MORQA (Yim et al., 2025a). The competition utilizes two primary multimodal datasets cantaining both image and textual data, i.e. WoundCare (Yim et al., 2025b), which focuses on wound management, and IIYI (Yim et al., 2024), which covers broader clinical cases. A series of questions were posed to three distinct models on the various cases proposed by the datasets, and the responses were then compared. The primary objective is to automatically evaluate the responses provided by the three systems, which are further subdivided into two subtasks: one in English (Subtask 1) and one in Chinese (Subtask 2). In the English subtask, systems are tasked with predicting six distinct criteria: disagreement flag, completeness, factual accuracy, relevance, writing style, and overall. With regard to the Chinese subtask, the evaluation is centered on two primary components: factual consistency and writing style. The scores ranged from 0 to 1 for each metric, where 0 indicates a terrible response (incorrect or/and harmful; may lead to grievous misinformation or dangerous outcomes; should never be shown to the patient), 0.5 indicates a response that is relevant and will not necessary cause harm but misses critical items or contains incorrect information (should not be shown to the patient before correction), and 1 indicates a complete and accurate response that can be safely presented to a patient. Note that disagreement flag was a binary choice.

The predictions were evaluated against the human scores using Pearson, Spearman, and Kendall's Tau correlations (Shaqiri et al., 2023).

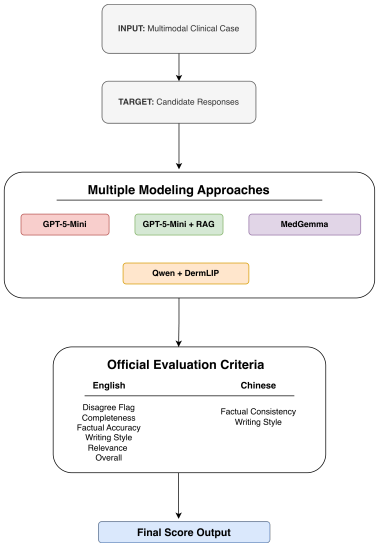


Figure 1: SloCal-Net workflow for multimodal evaluation of medical answers.

We evaluate several judging pipelines (Figure 1): generic reasoning with GPT-5-mini (Papineni et al., 2002) (zero-shot and RAG), specialized medical reasoning with MedGemma (Sellergren et al., 2025) and other state-of-the-art open source models (Section 5.5), and a local multimodal pipeline that combines DermLIP for visual features, KeyBERT (Grootendorst, 2020) and SBERT (Reimers and Gurevych, 2019) for semantic text processing, and Qwen-2.5 as the reasoning backbone.

3. Data and Resources

This study uses two publicly available multimodal clinical Visual Question Answering (VQA) benchmarks: **DermaVQA** (Yim et al., 2024) and **WoundcareVQA** (Yim et al., 2025b). Both datasets pair clinical images with natural-language questions and expert-written answers, but they differ in domain (dermatology vs. wound care), scale, and available metadata (Table 1). DermaVQA combines Chinese cases from IIYI with English cases from Reddit, and includes a subset with patient metadata (age/sex/Fitzpatrick skin type), while WoundcareVQA focuses on wound assessment with rich structured labels (e.g., anatomic location, wound type and stage, drainage, and infection status) and expert responses authored by US clinicians.

DermaVQA and WoundcareVQA are multilingual, image-grounded QA resources designed to reflect real patient–provider interactions (Yim et al., 2024, 2025b). DermaVQA centers on dermatology, com-

Feature	DermaVQA	WoundcareVQA
Medical Domain	Dermatology	Mainly Wound Care
Total Images	3,434	748
Total Cases/QA	1,488 question-answer pairs	477 cases (768 expert responses)
Data Sources	IIYI (Chinese) and Reddit (English)	Tieba and Zhidao (Chinese translated to English)
Clinical Metadata	Age, Sex, and Fitzpatrick Skin Type (Reddit subset only)	41 anatomic locations, 8 wound types, 6 thickness stages, tissue color, drainage, and infection
Expert Involvement	Medical translators (IIYI) and professional dermatologists (Reddit)	Practicing US medical doctors (Surgeon, ER Physician, ER Resident)

Table 1: Comparison between DermaVQA and WoundcareVQA datasets.

bining Chinese forum cases (IIYI) with English Reddit questions; WoundcareVQA targets wound assessment and triage, with expert responses and extensive structured metadata to support fine-grained clinical evaluation. A concise comparison of their domains, sizes, data sources, and metadata is provided in Table 1.

4. Tools Overview

The experimental setup combines multimodal reasoning models, domain-specific vision encoders, and semantic text-processing components. DermLIP (Yan et al., 2025) provides dermatology-oriented visual representations, GPT-5-mini (Wang et al., 2025) and MedGemma act as multimodal clinical judges, SBERT (Reimers and Gurevych, 2019) and KeyBERT (Grootendorst, 2020)support semantic alignment and keyword extraction, and Qwen models serve as multilingual reasoning backbones. Each component addresses a distinct stage of the evaluation pipeline, enabling controlled comparison across architectures.

- **GPT-5-mini** : A compact multimodal reasoning model used as the primary evaluation engine. It demonstrates strong performance on medical knowledge and multimodal QA benchmarks, supporting integrated interpretation of textual and visual clinical evidence.
- **Derm1M and DermLIP** : Derm1M is a large dermatology vision–language dataset aligned with clinical ontologies. DermLIP models trained on this resource enable fine-grained visual feature extraction and improved cross-modal retrieval in dermatological settings.
- **Sentence-BERT (SBERT)**: Produces sentence-level embeddings to estimate the semantic similarity between each candidate and a consensus clinical context—dynamically extracted from the pool of candidates via KeyBERT—to evaluate factual alignment.
- **KeyBERT** : Extracts clinically relevant keywords using contextual embeddings, aiding relevance and completeness assessment.
- **Qwen models** (Qwen Team, 2024; Team, 2024): Multimodal and multilingual transformer models used as reasoning backbones in the local pipeline, enabling structured inference across text and image inputs.

5. Methodology

Given the complexity of the WoundCare and IIYI datasets, we explored multiple judging strategies. We first tested RAG to provide external evidence to the judge at scoring time. We then enforced structured outputs so each model returned per-criterion scores in a consistent format. Next, we evaluated a fully local multimodal pipeline based on Qwen and DermLIP. Finally, we compared against MedGemma as a domain-specialized multimodal judge as well as general open-source models, such as Qwen3-VL and GLM-4.6V. Across all settings, the goal is to predict scores that reflect the judgment of clinical experts on a discrete scale $\{0, 0.5, 1.0\}$.

To evaluate the impact of these architectural choices, we configured our judging systems into five distinct experimental runs reported in Table 3:

- **GPT-5-mini WS**: A baseline multimodal judge using a web-search tool for real-time information retrieval without a static indexed corpus.
- **GPT-5-mini WS RAG**: The baseline system augmented with our 25-document clinical evidence corpus for grounded retrieval-augmented generation.
- **GPT-5-mini WS Struc**: The baseline system utilizing Pydantic-enforced schemas to generate rigid, multi-dimensional JSON scores.
- **GPT-5-mini WS RAG Struc**: A composite configuration combining both the static RAG corpus and structured output constraints.

- **DermLIP Qwen2**: Our fully local pipeline integrating specialized dermatological embeddings with a Qwen reasoning backbone.

Within these configurations, 'Web Search' (WS) refers to the model's native agentic ability to query live clinical resources, whereas 'RAG' specifically refers to grounding within our curated 279-page clinical document index.

5.1. Retrieval-Augmented Judging (RAG)

In the RAG setting, we augment the judge with external evidence so that scoring is explicitly grounded in verifiable background knowledge rather than relying solely on the model's parametric memory. We constructed an evidence corpus using an agentic web-search workflow (ChatGPT Deep Research) that (i) analyzed the official task website and evaluation protocol, and (ii) searched the literature for high-quality resources on multimodal medical reasoning, medical image interpretation (dermatology and wound care), and reliable evaluation methodologies for open-ended QA. The Deep Research prompt we used was:

> I need a comprehensive background knowledge for this competition https://sites.google.com/view/mediqa2026/mediqa-eval Analyse all the data from the website and search the web for resources that can the best ground a model, note that this mean both understanding medical images or how to interpret them and provided text. At the end provide 25 hyperlinks to the best knowledge to get a grounded understanding and how to best evaluate a statement. Be very scientific and precise. The link need have access to a pdf file, so double check it, since it will be use a additional context in a RAG.

5.2. Structured Output and Scoring

To assess an entire clinical encounter in a single multimodal inference step, we used prompts with Pydantic-enforced schemas. The model is required to generate: (i) **Visual Synthesis**, a detailed characterization of pathological findings in the provided images; (ii) **Integrated Reasoning**, a clinical chain-of-thought (CoT) that synthesizes visual evidence, case history, and expert references; and (iii) **Multi-Dimensional Scoring**, a structured JSON object containing discrete scores (0.0, 0.5, 1.0) for all required metrics (and, when applicable, a Disagree Flag). This structured format enables consistent aggregation and evaluation across judge models.

5.3. Qwen and DermLIP Pipeline

The proposed architecture, developed to test the dataset proposed for the competition, integrates visual and textual information through different modules, with the reasoning process guided by a Large

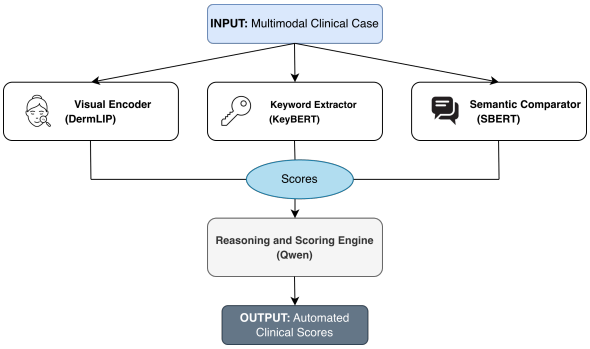


Figure 2: SloCal-Net workflow for Qwen and DermLIP.

Language Model. As shown in Figure reffig:workflow1, the evaluation process consists of three main stages: feature extraction, context integration, and final scoring. Given the specificity of the dermatological domain, clinical images are processed using DermLIP. This model, pre-trained on medical datasets, allows the extraction of visual embeddings that capture critical pathological details, overcoming the limitations of generalist visual models. The textual component is analysed along two lines:

- *Clinical Relevance:* We use KeyBERT to extract essential medical keywords from the entire pool of candidate responses, ensuring that key concepts are not overlooked.
- *Semantic Coherence:* We employ SBERT to calculate the vector similarity between the candidate's generated response and the reference, providing a quantitative measure of semantic alignment.

The system is configured to produce discrete scores in the range $\{0, 0.5, 1.0\}$ for different qualitative dimensions; only for Factual Accuracy are the values not discretised. For the English language subtask, the following are evaluated: Completeness, Factual Accuracy, Relevance, Writing Style, Overall Quality, and a Disagree Flag for clinical safety. For the Chinese language subtask, the system focuses on *Factual Consistency* and *Writing Style*.

For the Factual Accuracy metric, Qwen integrates the visual features provided by DermLIP into the decision-making process, allowing for clinical accuracy verification anchored to photographic evidence of the case. For the remaining metrics (such as Completeness, Relevance and Writing Style), Qwen acts as an informed zero-shot classifier, analysing the consistency between the keywords extracted by KeyBERT, the semantic similarity calculated by SBERT and the content of the candidate response to assign the final scores. The

Qwen model therefore constitutes the central reasoning unit of the pipeline and is used via a prompting technique; specifically, prompt-based assessment is reserved for qualitative aspects, where the model acts as a 'senior clinician' to assign discrete scores. In this context, Qwen evaluates the candidate's response by checking its clinical consistency and the use of professional terminology.

5.4. GPT-5-mini

We used GPT-5-mini as our primary multimodal judge, both in a zero-shot setup and with retrieval-augmented generation (RAG). In the zero-shot configuration, the model directly scored each response according to the official rubric criteria. In the RAG configuration, the judge was additionally grounded with a small set of retrieved external documents (produced via a web-search tool) intended to provide task-relevant medical context and reduce unsupported claims. This setup follows common patterns in recent evaluation competitions that leverage rubric-driven *LLM-as-a-judge* scoring and study grounding to improve factuality and calibration (Zheng et al., 2023; Anonymous, 2025; Asma Ben Abacha, 2026). To ensure reproducibility, the same vision-language model utilized as the primary scoring engine (GPT-5-mini) was also employed to generate the image-aware summaries used for constructing the retrieval queries.

5.5. Open-source models

We carried out a **post-competition** evaluation with the open-source multimodal models reported in Table 4:

- **GLM-4.6V-109B-Instruct** (Zhipu AI, 2025). A large-scale vision–language model from the GLM lineage designed for multimodal reasoning across images and documents. It supports instruction-following and visual grounding for tasks such as medical VQA and report interpretation.
- **Qwen3-VL-30B-Thinking** (Qwen Team, 2025). A reasoning-focused multimodal model optimized for structured inference over visual and textual inputs, building on the Qwen-VL framework for image understanding, grounding, and multimodal dialogue (Qwen Team, 2023). These capabilities enable diagnostic-style reasoning workflows combining imaging and clinical context.
- **Qwen3-VL-32B-Instruct** (Qwen Team, 2025). Instruction-tuned multimodal model for document parsing, visual grounding, and image-conditioned response generation, aligned with the general multimodal capabilities established in the Qwen-VL technical report (Qwen Team, 2023), including applications to medical image–text interpretation tasks.
- **MedGemma-27B-it** (Sellergren et al., 2025). A medical-domain vision–language model designed for clinical reasoning across radiology and pathology imagery. It supports clinical question answering and report-style generation.
- **MedGemma-1.5–4B-it** (Google, 2026). A compact multimodal clinical model optimized for deployment and fine-tuning; smaller medical vision–language backbones have been used for medical reasoning in resource-constrained settings.

We ran these models in the same experimental setup as our other judges: each model received the image, question, candidate answer, and rubric definition (and, in the RAG variant, the retrieved evidence context) and then produced per-criterion scores. These experiments were conducted after the official submission deadline and are therefore reported as supplementary analyses rather than leaderboard submissions. All open-source models were served locally on an HGX $8\times$H200 server via vllm; we used a single H200 GPU for all models except GLM-4.6V-109B-Instruct, which was run on two H200 GPUs.

6. Results and Discussion

6.1. Official shared-task results

Team	Score(en)	Score(zh)
MEDIQA	0.5551	0.2767
SUAT-BMI	0.5450	/
MedAware	0.5261	0.2505
SloCal-Net (Ours)	0.4656	0.2596
BDI	0.4392	0.2524
hgkai26	0.2444	/

Table 2: Official leaderboard results.

The official leaderboard performance (Table 2), measured as Pearson correlation between expert ratings and automatic scores for English and Chinese. Our primary submission (GPT-5-mini with web search and rubric-conditioned scoring) achieved correlations of 0.4656 (English) and 0.2596 (Chinese). These results suggest that general-purpose multimodal reasoning models can approximate expert judgment in rubric-driven grading tasks.

However, a performance gap remains relative to the top-ranked systems, particularly for English.

Strategy	RAG	Overall mean (en)	Overall mean (zh)	Latency (s)	Total Tokens
GPT-5-mini WS	No	**0.4898**	0.2555	30.64	3,155
GPT-5-mini WS RAG	Yes	0.4656	**0.2596**	26.93	3,805
GPT-5-mini WS Struc	No	0.3021	0.1649	28.47	3,092
GPT-5-mini WS RAG Struc	Yes	0.2436	0.1628	27.23	3,864
DermLIP Qwen2	No	0.2428	0.1678	16.10	614

Legend: WS = web-search tool; Struc = structured output.

Table 3: Submitted runs with latency and total tokens used.

This suggests that while strong reasoning capacity is necessary, calibration to task-specific rubrics and multimodal grounding is equally critical. The lower Chinese performance reflects broader challenges in multilingual alignment, including differences in rubric interpretation, linguistic ambiguity, and variability in clinical terminology. This performance gap is primarily driven by the English-centric bias in foundational medical pre-training data, which results in less reliable cross-modal grounding for Chinese clinical narratives. Furthermore, the inherent difficulty of translating highly nuanced clinical evaluation rubrics introduces subjective variance that disrupts the models' alignment with expert judgments.

6.2. Ablations: retrieval and output-structure effects

Looking at controlled ablations (Table 3), where we isolate the impact of retrieval and structured outputs on GPT-5-mini. We firstly observe, that retrieval provided a modest improvement for Chinese (0.2555 $\rightarrow$ 0.2596) but reduced performance for English (0.4898 $\rightarrow$ 0.4656). The deterioration in performance, in this case, may be linked to semantic noise and 'semantic bleeding', i.e. the inclusion of external documents may introduce irrelevant information that distracts the model.

A post-hoc analysis revealed that approximately 15% of retrieved passages contained 'semantic noise', medically accurate but contextually irrelevant protocols, which triggered 'semantic bleeding' and a 2.4% decrease in English Pearson correlation. Conversely, this same external context likely provided a necessary terminological bridge in the Chinese subtask, where retrieval noise was outweighed by the benefits of improved multilingual grounding. In multimodal contexts (text + image), this translates into "semantic bleeding"(Brown et al., 2025). The poor performance of specialised workflows (0.24) reflects a structural challenge in the field of dermatology linked, for example, to the so-called "granularity gap": models may be excellent at binary screening but fail dramatically when it comes to identifying specific subclasses (Yuceyalcin et al., 2026).

Furthermore, in the semantic space of the models, visually similar lesions tend to overlap. This overlap makes it almost impossible for workflows such as DermLip + Qwen to correctly separate pathologies that, although biologically different, appear similar to AI. The low score of the DermLip-based workflow is further justified by the technical limitations of CLIP-based models, such as text and token number limitations (Yan et al., 2025).

Our second observation is that structured output constraints reduced correlation for both languages. While structured scoring improves interpretability and reproducibility, it appears to over-constrain the judge and limit flexible reasoning across modalities. This might indicate trade-off between evaluation transparency and predictive alignment.

Latency and token consumption further reveal a practical trade-off: retrieval increased computational cost without consistent performance gains, suggesting that evidence selection and filtering are critical design factors.

6.3. Modular Ablation: Impact of DermLIP and KeyBERT-SBERT Context

To assess the contribution of each module used in the hybrid evaluation pipeline (Qwen + DermLIP + KeyBERT), we conducted a controlled ablation study on the validation set; the results for Factual Accuracy are shown in Table 5. The results show that the complete pipeline provides the most accurate alignment with expert evaluations. The most significant drop in performance occurred when the *Clinical Context* module was removed (**No-KeyBERT**). Similarly, the exclusion of specialised visual features (**No-DermLIP**) confirms that, whilst generic models such as Qwen possess strong reasoning capabilities, they lack the diagnostic sensitivity of expert-level dermatological latent spaces.

6.4. Post-competition evaluation of open-source judges

Post-competition experiments evaluated open-weight multimodal models under identical prompting conditions (Table 4). The strongest English

Model	RAG	Overall mean (en)	Overall mean (zh)	Latency (s)	Total Tokens
GLM-4.6V-109B-Instruct (Zhipu AI, 2025)	Yes	0.2716	0.2168	272.83	8,044
	No	0.2512	0.2617	179.51	3,155
MedGemma-27B-it (Sellergren et al., 2025)	Yes	0.2559	0.1789	44.57	5,568
	No	0.1729	0.1613	34.54	2,319
Qwen3-VL-30B-Thinking (Qwen Team, 2025)	Yes	0.2117	0.1533	45.55	6,289
	No	0.2045	0.1466	21.72	2,509
Qwen3-VL-32B-Instruct (Qwen Team, 2025)	Yes	0.1930	0.1616	47.27	6,137
	No	0.1875	0.1484	33.64	2,751
MedGemma-1.5-4b-it (Sellergren et al., 2025)	No	0.0899	0.0417	43.27	2,302
	Yes	0.0368	0.0821	46.03	5,543

Table 4: Post-competition analysis of open-source models.

Configuration	Pear. (r)	Spear. (ρ)	Kend. (τ)
Full (Qwen+DermLIP+KB)	**0.2188**	**0.2294**	**0.1810**
No-DermLIP (Vision Qwen)	0.1916	0.1934	0.1520
No-KeyBERT (No Context)	0.1365	0.1519	0.1200

Table 5: Ablation results for the Qwen+DermLIP scoring pipeline.

performance was observed for GLM-4.6V (0.27 Pearson), followed by MedGemma and Qwen3-VL variants. Performance differences across models correlate strongly with scale and multimodal pretraining depth.

The performance lags significantly proprietary models, such as GTP-5-mini, indicating limited multilingual calibration and weaker rubric grounding. Smaller models showed even larger degradation, suggesting insufficient capacity for joint reasoning across image, language, and evaluation criteria.

These findings indicate that open-source multimodal judges remain competitive for exploratory or resource-constrained settings but still lag behind proprietary models in reliability and cross-lingual robustness.

Future work will focus on adaptive evidence selection strategies, and improved multimodal grounding for multi-image and low-quality clinical inputs.

6.5. Limitations

The proposed evaluation framework presents three primary limitations. First, system efficacy relies heavily on the quality of the RAG knowledge base, where poor retrieval directly degrades the judge's diagnostic accuracy. Second, the multimodal judge exhibits high prompt sensitivity, meaning subtle lexical variations can significantly alter its interpretation of complex clinical rubrics. Finally, deploying state-of-the-art proprietary models like GPT-5-mini incurs substantial token costs, creating a practical financial impediment to scaling the framework for high-volume clinical analysis.

7. Conclusions

We presented SloCal-Net's participation in the MEDIQA-EVAL 2026 shared task, treating automated evaluation as rubric-conditioned multimodal judging with optional retrieval grounding. Our best system (GPT-5-mini with web search and structured scoring) achieved Pearson correlations of 0.466 (English) and 0.260 (Chinese) against expert ratings. Post-competition tests show that open-weight multimodal judges can approach the English leaderboard performance (up to 0.402 Pearson) but exhibit a larger drop on Chinese, suggesting that multilingual calibration and rubric interpretation remain key bottlenecks. Overall, retrieval and structured outputs can help, but they require careful prompt design to avoid over-constraining the judge. Future work will focus on multilingual rubric normalization, evidence selection, and robustness to multi-image cases and varying image quality.

8. Acknowledgements

This work was supported by European Union under Horizon Europe [grant number 101159018] and EuroHPC JU [grant number 101101903]; Slovenian Research Agency [grant number N3-0307, GC-0001]. Project Unveiling RNAbased Therapeutic Targets for PostTransplant Nephropathy (SINNephro RNA) PNRR Next Generation EU CN00000041 National Center for Gene Therapy and Drugs based on RNA Technology. National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.5, project 'RAISE – Robotics and AI for Socio-economic Empowerment' (ECS00000035) under the project 'Gestione e Ottimizzazione di Risorse Ospedaliere attraverso Analisi Dati, Logic Programming e Digital Twin (GOLD)', CUP H53C24000400006.

9. Bibliographical References

Anonymous. 2025. Healmqa: A healthcare multimodal question answering dataset for benchmarking large language models. OpenReview preprint (under review at ICLR 2026). Accessed: 2026-02-17.

Wen-wai Yim Asma Ben Abacha. 2026. Overview of the mediqa-eval 2026 shared task on evaluation metrics in medical multimodal question answering. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

Andrew Brown, Barry Devereux, and Muhammad Roman. 2025. Retrieval-augmented generation (rag): A systematic literature review of techniques, metrics, challenges, and future directions. *Big Data and Cognitive Computing*.

Google. 2026. MedGemma 1.5-4b-it (model card). Hugging Face model card. Accessed: 2026-02-17.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out (ACL 2004 Workshop)*, pages 74–81, Barcelona, Spain.

OpenAI. 2025. Deep research system card. System card. Accessed: 2026-02-17.

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. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint. ArXiv:2308.12966.

Qwen Team. 2024. Qwen2 technical report. arXiv preprint. ArXiv:2407.10671.

Qwen Team. 2025. Qwen3-vl technical report. arXiv preprint. ArXiv:2511.21631.

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. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763.

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 (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

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.

Mirlinda Shaqiri, Teuta ILJAZI, Lazim KAMBERI, and Rushadije RAMANI-HALILI. 2023. Differences between the correlation coefficients pearson, kendall and spearman. *Journal of Natural Sciences and Mathematics of UT*, 8(15-16):392–397.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. 2025. Capabilities of gpt-5 on multimodal medical reasoning. *arXiv preprint arXiv:2508.08224*.

Siyuan Yan, Ming Hu, Yiwen Jiang, Xieji Li, Hao Fei, Philipp Tschandl, Harald Kittler, and Zongyuan Ge. 2025. Derm1m: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12681–12690.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2025. Morqa: Benchmarking evaluation metrics for medical open-ended question answering. arXiv preprint. ArXiv:2509.12405.

Furkan Yuceyalcin, Abdurrahim Yilmaz, and Burak Temelkuran. 2026. A hierarchical benchmark of foundation models for dermatology. *arXiv preprint arXiv:2601.12382*.

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.

Zhipu AI. 2025. Glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable vision-language models. arXiv preprint. ArXiv:2507.01006.

10. Language Resource References

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2025a. MORQA: benchmarking evaluation metrics for medical open-ended question answering. volume abs/2509.12405.

Wen-wai Yim, Asma Ben Abacha, Robert Doerning, Chia-Yu Chen, Jiaying Xu, Anita Subbarao, Zixuan Yu, Fei Xia, M Kennedy Hall, and Meliha Yetisgen. 2025b. WoundcareVQA: A multilingual visual question answering benchmark dataset for wound care. *Journal of Biomedical Informatics*, 161:104639. Language resource / dataset paper.

Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2024. DermaVQA: A multilingual visual question answering dataset for dermatology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 209–219. Springer. Language resource / dataset paper.

A. Appendices

A.1. LLM Judge Prompts

This appendix details the exact system and user prompts used to instruct the LLMs for rubric-conditioned multimodal judging in both the English and Chinese subtasks.

A.1.1. English Evaluation Prompts

System Prompt:

```
You are a Senior Clinical Evaluator. The identified body part in the image is: {body_part}. Perform a rigorous EXTREMELY CONCISE assessment.
AVOID FLUFF and filler text.

Evaluation Protocol (MANDATORY):
1. **Visual Synthesis**: Concise, high-yield description of pathology and landmarks, focusing on the {body_part} region.
2. **Clinical Reasoning**: Brief, step-by-step analysis comparing Candidate vs. Gold Standard. Focus strictly Safety (red flags), Accuracy, and Actionability.
3. **Structured Scoring**: Assign scores (0.0, 0.5, 1.0). All justifications MUST be under 50 words.

Scoring Rubric:
- **completeness**: 1.0 (Full/Correct), 0.5 (Partial), 0.0 (Incorrect/Missing).
- **factual-accuracy**: 1.0 (Accurate), 0.5 (Minor errors), 0.0 (Wrong).
- **relevance**: 1.0 (Relevant), 0.5 (Partial), 0.0 (Irrelevant).
- **writing-style**: Professionalism, clarity, and conciseness.
- **overall**: Summary of utility and safety.
- **disagree_flag**: (0.0 or 1.0) 1.0 if advice contradicts expert.
```

User Prompt:

```
### CLINICAL CASE
Title: {title}
Content: {content}

### RAG CONTEXT
{rag_ctx}

### EXPERT REFERENCE (GOLD STANDARD)
{ref_text}

### CANDIDATE ANSWER TO EVALUATE
{candidate}

Provide your analysis and scores in structured JSON. You MUST provide a score for ALL metrics defined in the schema.
```

A.1.2. Chinese Evaluation Prompts

System Prompt:

```
您是资深医学专家评估员。图像中识别出的身体部位：您是资深医学专家评估员。图像中识别出的身体部位是：{body_part}。请进行严谨且极其简洁的评估，严禁废话和冗长描述。

评估流程：
视觉分析：结合已识别的部位（），简洁、高价值地1. **视觉分析**：结合已识别的部位（{body_part}），简洁、高价值地描述图像病理特征与解剖标志。
临床推理：简明扼要地对比专家标准，分析安全性（危险信号）与确。2. **临床推理**：简明扼要地对比专家标准，分析安全性（危险信号）与准确性。
结构化评分：为每个指标提供或的，理由必须在字以内。3. **结构化评分**：为每个指标提供 0.0, 0.5, 或 1.0 的评分，理由必须在 50 字以内。

评分细则：
- **factual-consistency-wgold**：
  - 1.0：完整且与至少一个参考回复事实相符。
  - 0.5：不完整但与至少一个参考回复事实相符。
  - 0.0：与所有参考回复事实不符。
- **writing-style**：基于专业性、清晰度和简洁度评分。
```

User Prompt:

```
### 患者案例
标题：{title}
内容：{content}

### RAG CONTEXT
{rag_ctx}

### EXPERT REFERENCE (GOLD STANDARD)
{ref_text}

### CANDIDATE ANSWER TO EVALUATE
{candidate}

Provide your analysis and scores in structured JSON. You MUST provide a score for ALL metrics defined in the schema.
```

MIDAS_SYNUR at MEDIQA-SYNUR 2026: A Prompting Study for Clinical Observation Extraction from Nurse Dictation Transcriptions

Swetha Krishna Sriram, Akshitaa Sahoo
MIDAS Lab, Indraprastha Institute of Information Technology Delhi (IIIT-Delhi)
New Delhi, India
kswetha036@gmail.com, akshitaa23068@iiitd.ac.in

Abstract

This paper describes MIDAS_SYNUR, a system developed for the MEDIQA-SYNUR task at ClinicalNLP 2026 on observation extraction from nurse dictations. The primary system adopts a single-prompt, field-rich few-shot strategy using GPT-5.2, jointly generating all schema fields in one structured output. Few-shot demonstrations are curated and grouped by value type, with five examples per type, promoting consistency across heterogeneous value distributions while leveraging global context to resolve cross-field dependencies. To analyze design trade-offs, this holistic strategy is compared against a field-wise decomposed prompting baseline, where each schema field is extracted independently using explicit positive and NULL demonstrations to improve absence detection and reduce cross-field interference. Zero-shot variants of both approaches are also evaluated to isolate the contribution of in-context examples. The results highlight inference-time prompting as a simple, reproducible, and competitive baseline for large-scale clinical observation extraction from conversational nurse dictations. **Keywords:** Prompt Engineering, Few-Shot Prompting, In-Context Learning, Structured Output

1. Introduction

Nurse dictation transcripts capture clinically important observations and care details, but they are expressed in conversational language and often lack explicit structure. Automatically extracting and normalizing these observations can reduce documentation burden and improve the completeness of patient records, yet it remains challenging due to variability in phrasing, implicit mentions, and frequent negations.

MEDIQA-SYNUR @ ClinicalNLP 2026 addresses this setting by requiring systems to populate a large set of predefined fields under a strict output schema. A practical challenge is deciding how to prompt large language models to cover many fields reliably: extracting fields independently can simplify NULL handling and reduce cross-field interference, while extracting all fields jointly can leverage global context and improve coherence.

In this work, we present MIDAS_SYNUR and focus on a comparative study of prompting strategies for observation extraction from nurse dictations. We evaluate two approaches using GPT-5.2: (i) *field-wise decomposed prompting*, where each field is extracted with a dedicated prompt and explicit positive/NULL demonstrations, and (ii) *single-prompt holistic extraction*, where all fields are predicted jointly in one structured prompt with value-type-grouped few-shot examples. We also include zero-shot variants to quantify the contribution of in-context examples, and discuss the resulting trade-offs in reliability, interference, and overall extraction quality.

2. Related Work

Clinical information extraction (IE) from unstructured text has been a long-standing research area in clinical natural language processing, driven by the need to transform free-text documentation into structured representations usable for downstream applications such as decision support, cohort identification, and workflow automation (Wang et al., 2018). Early clinical IE systems were predominantly rule-based, relying on hand-crafted patterns, lexicons, and domain ontologies to extract clinical concepts and events from electronic health records. Widely used systems such as MedLEE, MetaMap, and cTAKES demonstrated strong precision in constrained settings but suffered from limited scalability and portability across institutions due to variation in clinical language and documentation practices (Y. et al., 2012).

Subsequent work shifted toward statistical and machine learning approaches, including Conditional Random Fields and other feature-based models, which improved adaptability over purely rule-based systems. With the advent of deep learning, neural sequence models such as BiLSTMs and attention-based architectures further improved performance by learning distributed representations directly from data. Transformer-based models marked a significant advance, with domain-specific encoders such as BioBERT (Lee et al., 2019), ClinicalBERT (Alsentzer et al., 2019), and BlueBERT (Peng et al., 2019) achieving strong results across a range of clinical NLP tasks (Turchin et al., 2023), including entity recognition and relation extraction. Pre-training on biomedical or clinical corpora allows these models to better understand medical terminology, abbreviations, and multi-word expres-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 244–250
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

sions compared to general-purpose BERT. However, these approaches typically require substantial annotated data and computational resources for fine-tuning, which can limit their practicality in resource-constrained settings.

More recently, large language models (LLMs) have been explored for clinical information extraction using zero-shot or few-shot prompting. Prior work has shown that general-purpose LLMs can perform competitively on structured extraction tasks when guided by carefully designed prompts, sometimes without any domain-specific fine-tuning. Within the MEDIQA shared task series, several systems have demonstrated the effectiveness of prompt-based approaches (Mehta, 2025; Karim and Uzuner, 2025) for extracting structured medical orders from conversational transcripts, highlighting the role of schema constraints, in-context examples, and output validation in improving reliability.

Our work builds on this line of research by focusing on schema-constrained extraction for MEDIQA-SYNUR, emphasizing strict format compliance and evidence-backed extraction across a large number of heterogeneous clinical fields. Rather than pursuing domain-specific fine-tuning, we explore how prompt engineering and schema-aware inference with a general-purpose LLM can support scalable and reliable structured extraction from nurse dictation transcripts.

3. Task Description

The MEDIQA-SYNUR shared task (Michalopoulos et al., 2026) focuses on the extraction and normalization of structured clinical observations from nurse dictation transcripts. These transcripts are derived from spoken nurse documentation and contain detailed patient observations expressed in conversational and often unstructured language. The objective of the task is to automatically capture clinically salient information from nurse-patient conversations, normalize extracted observations, and map them to a large ontology of clinical concepts.

Given a nurse dictation transcript $D = \{u_i\}_{i=1}^{N}$, where each u_i represents an utterance segment within the dictation, systems are required to populate a predefined set of clinical fields specified by the official MEDIQA-SYNUR output schema. Each field corresponds to a particular clinical concept or observation and may require extracting values such as textual descriptions, categorical labels, numerical measurements, or normalized clinical indicators. Multiple fields may be populated from a single transcript, and not all fields are guaranteed to be present in every instance.

In addition to identifying relevant clinical information, systems must distinguish between clinically meaningful content and incidental or non-actionable language that commonly appears in spoken dictations. Submissions must also adhere to the provided schema by using the correct field identifiers and value types, and by handling missing information consistently (e.g., outputting NULL when a field is not supported by the transcript). Overall, the task emphasizes robust clinical language understanding together with reliable, structured prediction at scale.

4. Dataset Description

The MEDIQA-SYNUR dataset supports structured extraction of clinical observations from nurse dictation transcripts using a predefined schema. The task targets nursing documentation and aims to reduce documentation burden by automatically capturing clinically salient information from conversational nurse dictations. The dataset follows the schema design and annotation framework introduced in the SYNUR dataset (Corbeil et al., 2025).

The output schema defines 193 target clinical observation fields. Each field is specified by a unique identifier, descriptive name, and value type. Supported value types include `STRING`, `NUMERIC`, `SINGLE_SELECT`, and `MULTI_SELECT`. Categorical fields dominate the schema, with 130 `SINGLE_SELECT` and 12 `MULTI_SELECT` fields, and an average of approximately three enumerated values per categorical field.

The dataset is released with 122 training instances, 101 development instances, and 199 test instances. Across splits, nurse dictation transcripts are of comparable length, averaging approximately 180–185 words per instance. Despite the large schema size, individual transcripts populate only a small subset of fields: on average, 12–14 observations per instance, corresponding to less than 8% of the full schema. Fields not supported by the transcript are omitted from the output, resulting in sparse, variable-length structured annotations. Figure 1 shows the cumulative distribution of populated schema fields per instance across training, development, and test splits. The cumulative distribution of populated fields per instance is closely aligned across splits, indicating comparable structural density.

For evaluation, the test split contains only nurse dictation transcripts without gold annotations. Systems are required to generate complete, schema-compliant JSON outputs for each input instance, with gold labels withheld to enable blind evaluation.

5. Methodology

We frame MEDIQA-SYNUR as a schema-constrained extraction problem in which a large

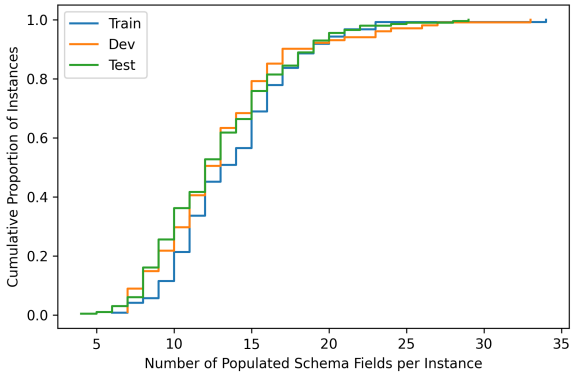


Figure 1: Cumulative distribution of the number of populated schema fields per instance across training, development, and test splits.

language model maps a nurse dictation transcript to a set of structured clinical observations defined by a fixed schema. We do not perform task-specific fine-tuning. Instead, we rely on prompt design, in-context examples, and post-generation validation to control extraction behavior, with particular emphasis on null handling and schema compliance.

All experiments are run using the OpenAI API with deterministic decoding (temperature = 0). The submitted system uses GPT-5.2, while additional models are evaluated for comparative analysis. The maximum output length is set to 16,384 tokens to avoid truncation. No constrained decoding or grammar-based enforcement is applied; instead, output validity is enforced through explicit prompt instructions and post-processing.

5.1. Prompt Design

We design structured prompt templates to control few shot sample usage, schema grounding, null handling, and output formatting. The prompts are organized into modular blocks that remain consistent across extraction strategies. A unified template is shown in Figure 2. Our prompt templates consist of the following blocks.

Block A: Core Instructions and Rules. This block contains the shared instructions used across all prompt variants. The model is instructed to extract information strictly from the provided transcript and not to infer, assume, or introduce information that is not explicitly stated. If a field is not supported by clear evidence in the transcript, it must be returned as `NULL` in the field-wise setting or omitted entirely in the single-prompt setting. The output must contain only the required fields in the specified format, with no additional explanation. The instructions allow limited clinical interpretation when mapping common clinical expressions to schema fields (e.g., device mentions), but guessing missing values is explicitly prohibited. When the evidence is unclear or indirect, the model is instructed not to extract the field. These rules aim to reduce unsupported extractions and ensure consistent handling of sparse or incomplete information.

For a subset of schema fields that are prone to systematic ambiguity, we include short, field-specific notes directly in the prompt. These notes are triggered by the field name and provide clarifications that distinguish closely related clinical concepts (e.g., bladder scan volume versus urinary output, suprapubic region versus abdomen) or require explicit mention of a specific assessment or checklist. This mechanism is used sparingly and only for fields where preliminary analysis showed consistent false positives without such clarification. The complete set of notes is shown in Figure 3. Notes were manually constructed based on error inspection on the development set.

Block B: Schema Exposure. The schema definition is injected directly into the prompt. For each field, the following metadata is provided: field identifier, field name, value type (`NUMERIC`, `SINGLE_SELECT`, `MULTI_SELECT`, `STRING`), and allowed enumeration values when applicable. This constrains the output space through explicit instruction rather than decoding constraints.

Block C: Demonstrations (Few-Shot Examples). Few-shot variants include minimal demonstrations. In both settings, demonstrations are drawn exclusively from the training set. For the field-wise setting, examples are selected as the first positive instance and first NULL instance found for each field when iterating over training examples in order. For the single-prompt setting, up to five positive examples per value type are collected by iterating through schema fields in order and taking the first matching training instance; once the quota for a type is reached, further examples of that type are skipped. In both cases, the same fixed set of demonstrations is used for all test instances, i.e., examples are not selected dynamically per input. Demonstrations are explicitly marked as format guidance and are not to be used as evidence for the target instance.

Block D: Output Specification and Target Instance. The final block contains the target transcript and the required output format. The expected structure differs across strategies, as described below.

```
BLOCK A: CORE RULES
------------------
- Use ONLY the TARGET TRANSCRIPT as
evidence.
- Do NOT infer, guess, or paraphrase.
- Do NOT use examples as evidence for
answers.
- If not explicitly mentioned:
  * Field-wise: output NULL
  * Single-prompt: omit field
- ...
- <May include specific notes for
certain fields>

BLOCK B: SCHEMA EXPOSURE
-----------------------
For each field:
- Field ID
- Field name
- Value type
- Allowed values (if any)

Type constraint example:
NUMERIC:
- Extract digits only.
- No units or text.

BLOCK C: DEMONSTRATIONS
----------------------
(Few-shot variants only)

Positive example:
Transcript: "..."
Output: <value>

NULL example:
Transcript: "..."
Output: NULL

BLOCK D: OUTPUT
----------------------
TARGET TRANSCRIPT:
"<text>"

OUTPUT FORMAT:
Field-wise: ANSWER: <value | NULL>
Single-prompt: valid JSON
```

Figure 2: Illustrative prompt template with block-wise structure.

5.1.1. Field-Wise Decomposed Extraction

In the field-wise strategy, each schema field is extracted independently using a separate model call. Block B contains only the metadata for the target field. The required output is a single value or `NULL`. Type-specific constraints (e.g., numeric-only output for `NUMERIC` fields, strict enumeration matching for categorical fields) are defined in the schema exposure block. This strategy isolates extraction decisions but requires one model call per schema field, adding up to `~190` calls per transcript.

5.1.2. Single-Prompt Holistic Extraction

In the single-prompt strategy, the full schema is provided in Block B. The model is instructed to return a valid JSON containing only fields explicitly supported by the transcript. Unsupported fields must not appear in the output. Unlike the field-wise formulation, this approach allows cross-field contextual reasoning within a single generation while reducing inference cost to one model call per transcript.

5.1.3. Zero-Shot Variants

To evaluate the contribution of demonstrations, we remove Block C while keeping Blocks A, B, and D unchanged. This isolates the effect of in-context examples without altering rule enforcement or schema exposure.

6. Results and Discussion

Our submitted system initially achieved an F1 of 0.773 on the MEDIQA-SYNUR leaderboard, ranking 7th. The top-ranked system obtained an F1 of 0.814. The submitted configuration uses GPT-5.2 in a single-prompt holistic extraction setting with type-wise few-shot demonstrations.

Post-competition, we identified inconsistencies in predicted field IDs caused by leading zeros (e.g., `02` instead of `2`). These arose because the model occasionally reproduced field IDs with zero-padding, which caused misalignments during evaluation since the official schema uses unpadded integers. We applied a one-time post-processing step that strips leading zeros from all predicted field IDs before evaluation. This normalization improved our submitted F1 from 0.773 to 0.82. All results reported below use this corrected output. No other modifications to predictions were made.

6.1. Overall Performance

Table 1 presents a comparative analysis of prompting strategies explored during development, including single-prompt and field-wise extraction settings under both few-shot and zero-shot conditions for GPT-5.1 and GPT-5.2. All reported results are computed on the official test set, for which gold annotations were released after the competition concluded.

We find that GPT-5.2 outperforms GPT-5.1 across all prompting variants, with F1 gains of $\sim$0.25–0.30. Recall exceeds precision for GPT-5.1

```
BLOCK A: FIELD-SPECIFIC NOTES
------------------------------
Applied sparingly to fields with
systematic false positives:

[Voiding function]
NOTE: Extract a value ONLY if the
term "Voiding function" is explicitly
used in the target transcript.

[Speech content]
NOTE: Refers to the nature of the
patient's communication or speech.
Extract ONLY if such nature of
communication/speech is explicitly
mentioned.

[Broset violence checklist fields]
NOTE: A specific checklist that must
be noted/performed by the nurse.
Extract (Yes/No) ONLY if the
transcript EXPLICITLY mentions the
field name, i.e., both the checklist
and the attribute together.

[Bladder scan volume]
NOTE: Bladder scan volume is different
from urinary output/volume. Look for
explicit mentions of Bladder scan
volume only.

[Fall risk identification]
NOTE: Refers to a patient's risk of
falls or conclusions from a fall risk
assessment. Extract the level of risk
ONLY if explicitly mentioned.

[Patient safety]
NOTE: Answer Yes ONLY if measures for
patient safety are explicitly stated
as taken. Answer No ONLY if explicitly
stated as not taken/not required.
Answer NULL if not explicitly mentioned.

[Patient identification]
NOTE: Refers to age and gender only.
Age must be an exact number (e.g.,
"X years old"); ignore descriptors
like "elderly" or "young". Extract
ONLY if both age and gender appear
together (e.g., "X-year-old female").

[Abdomen fields]
NOTE: Suprapubic region is NOT the
same as abdomen.

These notes were manually constructed
from development set error analysis.
```

Figure 3: Field-specific clarification notes included in Block A for fields prone to systematic ambiguity.

in all settings, suggesting a bias toward extracting potentially relevant fields even when evidence is weak. In addition to task-specific factors, the performance gap between GPT-5.1 and GPT-5.2 is consistent with generally reported improvements in the latter model. On this task, GPT-5.2 achieves higher precision and recall across all prompting configurations, suggesting more reliable extraction overall. GPT-5.1 underperforms on both dimensions, with particularly low precision, indicating a tendency to predict fields even when transcript evidence is weak. These gains likely reflect broader improvements in instruction-following and structured output generation in GPT-5.2, though the precise reasons remain difficult to isolate from prompt-level evaluation alone.

Few-shot prompting is seen to consistently improve precision and recall. Field-wise prompting offers modest gains in recall, with an advantage that likely stems from clearer task scoping. However, the margin is small, indicating that well-structured unified prompts can achieve comparable performance at a much lower inference cost.

Model	Approach	P	R	F1
GPT-5.1	Single-prompt (few-shot)	0.47	0.65	0.54
GPT-5.1	Field-wise (few-shot)	0.43	0.68	0.53
GPT-5.1	Single-prompt (zero-shot)	0.41	0.64	0.50
GPT-5.1	Field-wise (zero-shot)	0.41	0.66	0.51
GPT-5.2	Single-prompt (few-shot)	**0.82**	**0.81**	**0.82**
GPT-5.2	Field-wise (few-shot)	0.82	0.83	**0.83**
GPT-5.2	Single-prompt (zero-shot)	0.77	0.75	0.76
GPT-5.2	Field-wise (zero-shot)	0.77	0.75	0.76

Table 1: MEDIQA-SYNUR results (precision, recall, F1).

6.2. Type-Level Performance

Tables 2 and 3 present type-level analysis for GPT-5.1 and GPT-5.2 in the single-prompt holistic extraction setting.

GPT-5.2 shows substantial gains across numeric and single-select fields, driven by improved unit parsing, schema adherence, and reduced false positives. Multi-select fields improve moderately, while string fields remain challenging due to sparsity and ambiguous textual cues. Detailed analysis of string fields reveals that many failures are

Field Type	Precision	Recall	F1
MULTI_SELECT	0.680	0.708	0.693
NUMERIC	0.473	0.798	0.594
SINGLE_SELECT	0.466	0.641	0.539
STRING	0.020	0.051	0.029

Table 2: Type-level performance for GPT-5.1 (single-prompt few-shot).

Field Type	Precision	Recall	F1
NUMERIC	0.919	0.965	0.941
MULTI_SELECT	0.939	0.817	0.874
SINGLE_SELECT	0.846	0.832	0.839
STRING	0.021	0.031	0.025

Table 3: Type-level performance for GPT-5.2 (single-prompt few-shot).

caused by exact-match evaluation, case or whitespace differences, minor formatting variations (e.g., “75 mL” vs “75mL”), and predictions that include extra descriptive text. Moreover, the model exhibits a tendency to overgenerate text, producing verbose outputs that partially match gold values but are penalized under strict matching. These insights suggest that STRING field evaluation may benefit from partial-match or semantic similarity metrics to better capture meaningful predictions.

6.3. Field-Level Analysis

Field-level performance was analyzed for GPT-5.2 and compared to GPT-5.1 for the single-prompt holistic extraction setting. We highlight trends using a bar plot of the top 5 fields with the largest F1 improvements. On average, GPT-5.2 improved field-wise F1 scores by 0.271 over GPT-5.1.

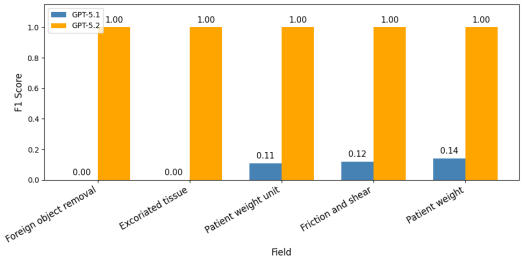


Figure 4: Top 5 fields showing the largest F1 score improvement from GPT-5.1 to GPT-5.2.

6.4. Error Analysis

We identify four main failure modes from examining test set predictions.

STRING field verbosity. The most consistent failure mode across both models is on STRING fields (F1: 0.025–0.029) with severe degradation in performance. The model frequently generates verbose outputs that partially match gold values but are penalized under strict evaluation. These errors suggest that STRING field evaluation would benefit from semantic similarity scoring rather than string equality.

False positives from cross-field confusion. A recurring error in both field-wise and single-prompt settings is the prediction of semantically related but distinct fields. For instance, urinary output values are occasionally extracted for the Bladder scan volume field, and abdomen-related observations are predicted for suprapubic fields. These confusions motivated the field-specific clarification notes described in Section 5.1. Their persistence in the results suggests that even explicit disambiguation notes do not fully resolve ambiguity in cases where clinical phrasing is inherently overlapping.

NULL prediction errors. GPT-5.1 shows a strong bias toward predicting fields even when transcript evidence is weak, reflected in its consistently low precision across all settings. GPT-5.2 largely corrects this, though occasional false positives remain, particularly for SINGLE_SELECT fields where a plausible enumerated value exists even without direct transcript support.

Schema compliance failures. A small number of predictions contain formatting inconsistencies such as zero-padded field IDs, extra whitespace, or unit strings appended to NUMERIC values (e.g., `75 bpm` instead of `75`). Padding inconsistencies are addressed by post-processing but the other issues remain.

7. Conclusion

In this work, we presented **MIDAS_SYNUR**, a prompting-based system for schema-constrained extraction of clinical observations from nurse dictation transcripts. Our experiments show that **GPT-5.2 with few-shot examples** achieves the strongest performance, consistently outperforming GPT-5.1 across all evaluation metrics.

We find that the single-prompt holistic extraction approach is preferable to field-wise decomposed extraction from a deployment standpoint. While field-wise few-shot extraction achieves a marginally higher F1 (0.83 vs. 0.82), the single-prompt approach requires only one inference call per transcript compared to approximately 190 for field-wise extraction, resulting in substantially reduced computation time and lower token usage. Given this cost difference, the single-prompt approach represents a more practical choice when the marginal F1 gap does not justify the added inference overhead.

Beyond overall performance, we provide a detailed analysis of predictions at both the type and field level. `STRING` fields remain the primary weakness, with errors largely due to sparse mentions, minor formatting inconsistencies, and strict exact-match evaluation criteria. These insights suggest potential avenues for future improvement, such as semantic similarity scoring or relaxed matching strategies for textual fields.

Overall, our study demonstrates that carefully designed prompt strategies, combined with few-shot demonstrations, provide a simple, reproducible, and effective framework for large-scale structured extraction from conversational clinical documentation.

8. Ethics Statement

This work benchmarks large language models (GPT-5.1 and GPT-5.2) in a controlled evaluation setting. As these models are pretrained on large-scale data, they may reflect and amplify societal biases related to language, culture, or demographics. Accordingly, performance differences should not be interpreted as purely objective measures of capability.We note that these models are proprietary and subject to change, which may limit full reproducibility. We mitigate this by clearly documenting prompts and evaluation settings. Finally, this work is intended for research purposes only. The evaluated models may produce incorrect or misleading outputs, and should not be deployed in high-stakes settings without appropriate safeguards and human oversight.

9. 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.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Suzhou (China). Association for Computational Linguistics.

A. H. M. Rezaul Karim and Özlem Uzuner. 2025. Masonnlp at mediqa-oe 2025: Assessing large language models for structured medical order extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 57–67. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Parth Mehta. 2025. Pnlp at mediqa-oe 2025: A zero-shot prompting strategy with gemini for medical order extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 75–83. Association for Computational Linguistics.

George Michalopoulos, Jean-Philippe Corbeil, Cari Bader, Nate Bodenstab, and Asma Ben Abacha. 2026. Overview of the mediqa-synur 2026 shared task on observation extraction from nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Alexander Turchin, Stanislav Masharsky, and Marinka Zitnik. 2023. Comparison of bert implementations for natural language processing of narrative medical documents. *Informatics in Medicine Unlocked*, 36:101139.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.

Wu Y., Denny J.C., Rosenbloom S.T., Miller R.A., Giuse D.A., and Xu H. 2012. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. In *AMIA Annual Symposium Proceedings*, pages 997–1003. PMID: 23304386.

AnotherOne at MEDIQA-SYNUR 2026: Detect, Extract, Normalize - Knowledge-Grounded LLM Pipeline for Clinical Observation Extraction

Jerrin John Thomas, Parameswari Krishnamurthy
International Institute of Information Technology, Hyderabad
jerrin.thomas@research.iiit.ac.in, param.krishna@iiit.ac.in

Abstract

We present a system for the MEDIQA-SYNUR 2026 shared task on extracting structured clinical observations from nurse dictation transcripts. The transcripts contain spoken-style clinical language with disfluencies, filler words, and hesitations. Our approach is a four-stage LLM inference pipeline preceded by an offline knowledge enhancement step: (1) knowledge-enhanced concept detection using medical domain clustering, (2) evidence-grounded value extraction, (3) schema-constrained value normalization, and (4) deterministic post-processing with fuzzy matching and unit pairing. In the offline step, we use the task ontology and training examples to generate per-concept clinical definitions and extraction rules, and group the 193 concepts into 19 non-exclusive medical domain clusters. These are injected into all downstream prompts as domain priors. All LLM stages use gpt-oss-120b with structured JSON output and chain-of-thought reasoning. The task requires exact matching on concept ID and value pairs across a 193-concept ontology, making precision particularly challenging. We iteratively refine concept definitions and prompt guidelines based on error analysis of the training data. Our system achieves an F1 score of 0.806 on the test set.

Keywords: clinical NLP, information extraction, nursing documentation, large language models, ontology-constrained extraction

1. Introduction

Nurses spend a disproportionate amount of time on electronic health record (EHR) documentation rather than direct patient care (Tierney et al., 2024). Automating the conversion of spoken nursing assessments into structured flowsheet entries can substantially reduce this documentation burden (Quiroz et al., 2019).

The MEDIQA-SYNUR 2026 shared task (Ben Abacha et al., 2026) addresses this problem directly: given a raw nurse dictation transcript containing filler words, hesitations, and disfluencies, systems must extract all clinical observations, map each to one of 193 predefined concepts in a structured ontology, and normalize values to the correct type. The task builds on the SYNUR dataset introduced by Corbeil et al. (2025), who demonstrated that LLMs can structure clinical speech transcripts but that the task remains challenging due to the strict exact-match evaluation criterion.

Several factors make this task difficult: (a) noisy spoken-style input with disfluencies and filler words; (b) a 193-concept ontology with strict exact-match evaluation on concept ID and value pairs; (c) limited training data (122 samples); (d) high annotation density, with approximately 14 observations per transcript; (e) negation handling where "no edema" must map to Edema: No edema; and (f) linking each numeric measurement to its corresponding unit concept.

Rather than a single end-to-end LLM call, we decompose the task into specialized stages - detect, extract, normalize, validate, each operating within constrained output spaces. We cluster concepts by medical domain, since transcripts documenting one concept in a clinical area typically document others from the same area (e.g., a respiratory assessment typically covers breath sounds, oxygen saturation, and cough description together).

Our contributions are: (1) a decomposed pipeline separating concept detection, value extraction, and normalization into distinct stages; (2) medical-domain concept clustering that exploits co-occurrence patterns in clinical assessments; (3) an offline knowledge enhancement step generating per-concept definitions and extraction rules as LLM priors; and (4) an iterative refinement process using training-set error analysis and model uncertainty signals to improve concept definitions, extraction rules, and prompt guidelines.

2. Related Work

Clinical NLP and Information Extraction. Early work in clinical information extraction includes the i2b2 challenges on concept, assertion, and relation extraction from clinical text (Uzuner et al., 2011). More recently, LLM-based extraction works have shown that providing concept definitions in prompts improves biomedical NER by 0.15 F1 score points (Hu et al., 2024b), and that prompt engineering alone can make LLMs competitive with fine-tuned models for clinical NER (Hu et al., 2024a).

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 251–256
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

MEDIQA Shared Tasks. The MEDIQA series has addressed clinical question answering (Ben Abacha et al., 2019), dialogue summarization (Ben Abacha et al., 2023), medical error detection (Ben Abacha et al., 2024), and medical order extraction (Ben Abacha et al., 2025). MEDIQA-SYNUR extends this to nursing observation extraction, building on the SYNUR dataset introduced alongside the closely related SIMORD dataset for medical order extraction (Corbeil et al., 2025).

Processing Clinical Dictation. Nurse dictation transcripts contain disfluencies, filler words, and informal language that pose challenges distinct from polished clinical notes. Corbeil et al. (2025) evaluated both open- and closed-weight LLMs on the SYNUR task, finding that larger models handle spoken-style clinical text more reliably. The documentation burden problem (Tierney et al., 2024; Quiroz et al., 2019) has driven work on structuring clinical speech.

Chain-of-Thought Reasoning. Wei et al. (2022) demonstrated that chain-of-thought prompting improves complex reasoning in LLMs. We adopt this by requiring explicit reasoning fields in every LLM output, forcing the model to articulate extraction logic before producing structured answers.

Negation in Clinical Text. Negation detection is critical in clinical IE. For example, "no facial droop" must map to Facial droop: no facial droop. The ConText algorithm (Harkema et al., 2009) introduced patterns for negation, experiencer, and temporal status. We incorporate these principles as prompt guidelines rather than rule-based preprocessing.

3. Task Description

The MEDIQA-SYNUR 2026 shared task (Ben Abacha et al., 2026) requires systems to extract structured clinical observations from synthetic nurse dictation transcripts. Given a raw transcript containing spoken-style clinical language with filler words, hesitations, and disfluencies, the system must identify which clinical concepts are documented, extract their values, and normalize the output to match a predefined ontology schema.

Each observation specifies a concept ID, concept name, value type, and value. The ontology defines 193 clinical observation concepts spanning vital signs, respiratory, cardiac, neurological, gastrointestinal, genitourinary, musculoskeletal, skin, safety, and behavioral domains.

Split	Samples	Gold Obs.	Avg. Obs.
Train	122	1,774	14.5
Dev	101	1,390	13.8
Test	199	2,721	13.7

Table 1: Dataset statistics. Avg. Obs. is the mean number of gold observations per transcript.

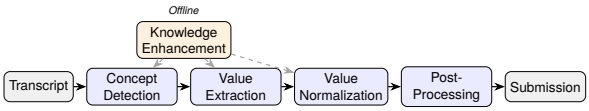


Figure 1: System architecture. Dashed arrows indicate domain knowledge (concept definitions, extraction rules, and cluster assignments) injected into LLM prompts. Post-processing is deterministic.

3.1. Dataset

The MEDIQA-SYNUR 2026 dataset (Ben Abacha et al., 2026; Corbeil et al., 2025) consists of synthetic nurse dictation transcripts in English with gold-standard structured observations. Table 1 summarizes the dataset statistics.

The 193 concepts span four value types: approximately 130 SINGLE_SELECT concepts with enumerated options, 25 STRING concepts requiring free-text values, 20 NUMERIC concepts with associated unit observations, and 15 MULTI_SELECT concepts. An additional 15 unit observation concepts accompany the numeric ones.

Evaluation. Performance is measured by instance-level Precision, Recall, and F1 score with exact matching on concept ID and value pairs. For MULTI_SELECT, values are unrolled into individual ID-value pairs. Numeric predictions are cast to a common numeric type before comparison.

4. System Description

4.1. Overview

Our system is a modular pipeline with four inference stages preceded by an offline knowledge enhancement step. Figure 1 illustrates the architecture. Each stage operates on the output of the previous one, progressively refining the extraction from broad concept detection to schema-compliant normalized observations. All LLM stages output structured JSON with chain-of-thought reasoning.

4.2. Knowledge Enhancement (Offline)

Two offline LLM-assisted steps prepare the domain knowledge injected into all downstream prompts.

Concept Definitions and Extraction Rules. For each of the 193 schema concepts, we prompt the LLM with the ontology and training examples to generate: (a) a clinical definition, and (b) 3-5 general extraction rules. For example:

- **Oxygen saturation** (NUMERIC): *Definition*: "Percentage of hemoglobin binding sites occupied by oxygen, typically measured via pulse oximetry." *Rules*: "Extract only when explicitly documented in the transcript", "Extract the numeric value only; extract the unit as a separate observation", "Do not infer a value from qualitative descriptions such as 'hypoxemic' "
- **Nausea** (SINGLE_SELECT, Yes/No): *Definition*: "Presence or absence of the subjective sensation of an urge to vomit." *Rules*: "Extract the affirmative option when nausea is stated as present", "Extract the negative option only when explicitly denied (e.g., 'no nausea', 'denies nausea')", "Do not assume absence; a negative value requires explicit negation"

Rules focus on when to extract vs. not extract, and are deliberately general. Normalization is handled in a later stage. The prompt instructs the LLM to produce a structured mapping of definitions and rules for each concept.

Medical Domain Clustering. The 193 concepts are grouped into 19 non-exclusive clusters based on medical domain knowledge and co-occurrence patterns in the training data (Table 2). The guiding principle is that a transcript documenting one member of a cluster is likely to document others from the same cluster, reflecting the structure of nursing assessments (e.g., a respiratory assessment typically documents breath sounds, oxygen saturation, cough description, and oxygen delivery device together). Clusters are strictly medical categories, not output-type groups, and their union covers all 193 concepts. Cluster boundaries were determined by grouping concepts that co-occur in the same clinical assessment context (e.g., cardiovascular concepts that appear together in a cardiac exam), then splitting overly large groups to keep per-cluster prompts focused.

The clustering is configurable: both the number of clusters and their composition can be redefined. Our configuration of 19 clusters balances prompt focus with full coverage.

4.3. Stage 1: Concept Detection

The concept detection stage identifies which of the 193 concepts are explicitly mentioned in each transcript. For each transcript-cluster pair, a single LLM call determines which concepts within that cluster are present.

Cluster	Size
Misc. GI & Body	23
Cardiovascular & Hemodynamics	18
Misc. Functional Mobility	16
Measurements & Units	15
Genitourinary	14
Neurologic & Mental Status	14
Misc. Vitals & Respiratory	13
Misc. Neuro & Sensory	12
Respiratory	11
Scores, Scales & Assessments	11
Nutrition & Hydration	11
Devices, Lines & Tubes	10
Gastrointestinal	10
Skin, Wound & Pressure Injury	9
Behavioral & Violence Risk	7
Mobility, Falls & Functional	7
Nursing Care & Environment	7
Admin. & Education	4
Pain	4

Table 2: The 19 medical domain clusters. Size indicates the number of concepts assigned to each cluster. Clusters are non-exclusive: a concept may appear in multiple clusters.

The prompt provides guidelines for explicit-only detection, negation-as-present (e.g., "denies nausea" means the Nausea concept is present with a negative value), and synonym awareness (e.g., "SpO2" = "Oxygen saturation"). For each concept in the cluster, the prompt includes: the concept ID, name, value type, definition, rules, and valid options for SELECT types.

The LLM outputs structured reasoning, detected concepts with evidence quotes and confidence scores, and a list of uncertainties. Responses are aggregated per transcript and deduplicated by concept ID across clusters.

4.4. Stage 2: Value Extraction

A single LLM call per transcript extracts raw values for all concepts detected in Stage 1. Evidence quotes from the detection stage are passed forward to ground the extraction.

The prompt provides type-specific extraction rules:

- NUMERIC: Extract the number only (e.g., "85" not "85%"); the unit is extracted as a separate observation using a predefined mapping from numeric concepts to their unit concepts.
- SINGLE_SELECT: Match one of the valid options.
- MULTI_SELECT: Return a list of matching options.
- STRING: Extract the minimal clinical phrase, removing filler words.

Explicit negation handling guidelines instruct the LLM to interpret "no jugular venous distention" as

JVD: “Absent” and “denies pain” as a negative indicator. As in Stage 1, structured reasoning and evidence are output alongside each extraction.

4.5. Stage 3: Value Normalization

A single LLM call per transcript normalizes all extracted values to match the schema exactly. The prompt provides strict normalization rules:

- SINGLE_SELECT: Output must exactly match one valid option (case-sensitive).
- MULTI_SELECT: Output a JSON array where each element matches a valid option.
- NUMERIC: Output a bare number (integer or float), stripping units.
- STRING: Clean the phrase by removing filler words.

For each concept, the full list of valid options is provided in the prompt. The LLM must also produce separate unit observations for numeric values and explain each normalization decision.

4.6. Stage 4: Post-Processing

A deterministic post-processing stage applies five operations:

Schema Validation. Each observation is checked against the ontology schema. Unknown concept IDs are rejected and value types are enforced.

Fuzzy Matching. For SELECT types, exact matching is attempted first. If no exact match is found, a fuzzy matching score is computed against all valid options, accepting the closest match above a similarity threshold. This recovers near-miss variants produced by the LLM.

Numeric Coercion. Unit suffixes are stripped (e.g., “%”, “bpm”, “mmHg”), blood pressure formats are preserved, and numbers embedded in text are extracted via regular expressions.

Deduplication. Duplicate ID-value pairs are removed.

Unit Pairing. A predefined mapping associates each numeric concept with its unit concept (e.g., Oxygen saturation → Oxygen saturation unit). If a numeric observation is present but its unit is missing, a default unit is inserted (e.g., “%” for O_2 saturation, “bpm” for heart rate), matched against the unit concept’s valid options.

System	Split	P	R	F1
Initial	Train	0.656	0.855	0.743
	Dev	0.641	0.863	0.736
	Test	0.645	0.886	0.747
Refined	Dev	0.761	0.839	0.798
	Test	0.755	0.865	0.806

Table 3: System performance across splits. Initial is the submitted system; Refined incorporates domain knowledge improvements from training-set error analysis (Section 5.3).

	Initial			Refined		
Type	**P**	**R**	**F1**	**P**	**R**	**F1**
SINGLE_SEL.	0.716	0.898	0.797	0.815	0.868	0.841
MULTI_SEL.	0.846	0.945	0.893	0.859	0.959	0.906
NUMERIC	0.742	0.984	0.846	0.929	0.965	0.947
STRING	0.020	0.102	0.033	0.020	0.061	0.030
Overall	**0.645**	**0.886**	**0.747**	**0.755**	**0.865**	**0.806**

Table 4: Per-type performance on the test set. Initial is the submitted system; Refined incorporates domain knowledge improvements.

4.7. Infrastructure

All LLM stages use gpt-oss-120b with structured JSON output. Calls are batched across transcripts for efficiency. gpt-oss-120b is a cost-effective model compared to current frontier models.

5. Results and Discussion

5.1. Main Results

Table 3 presents performance across data splits for both the initial submitted system and the refined system (Section 5.3).[1]

The initial system scores are consistent across splits, with F1 ranging from 0.736 on dev to 0.747 on test. The refined system improves F1 to 0.798 on dev and 0.806 on test, with the gains concentrated in precision.

5.2. Analysis by Value Type

Table 4 breaks down test set performance by value type for both systems.

MULTI_SELECT achieves the highest F1 (0.893 initial, 0.906 refined) because the constrained option lists reduce ambiguity. NUMERIC shows near-perfect recall in the initial system (0.984); the refined system trades a small amount of recall for a large precision gain (0.742→0.929), primarily by

[1]The official submission scored F1=0.703 due to a concept ID formatting mismatch (leading zeros). All scores in this paper use normalized IDs.

resolving the confusion between Oxygen saturation and Pulse oximetry concepts. SINGLE_SELECT, the largest category, improves from 0.797 to 0.841.

STRING observations remain the clear weakness, with F1 ~0.03 in both systems. The LLM generates plausible clinical phrases that are semantically correct but lexically different from the gold standard, and exact-match evaluation penalizes these mismatches.

5.3. Knowledge Refinement

Each LLM stage outputs an uncertainty list alongside its predictions, enabling systematic error analysis. We ran the initial system on the training set (F1=0.743) and analyzed both the error patterns and the uncertainty signals from each module. This analysis used only the training set; no development or test data informed the refinements.

The training-set analysis identified five systematic issues: (1) *concept confusion* between Oxygen saturation and Pulse oximetry, where the model extracted the same SpO2 reading under both concept IDs; (2) *over-inference of behavioral concepts*, particularly the Broset Violence Checklist confusion item being triggered by any mention of "confusion" rather than requiring an explicit BVC reference; (3) *STRING hallucination* for concepts like Patient identification and Safety equipment, where the model generated plausible observations not present in the gold annotations; (4) *unit hallucination* where unit observations were extracted without a corresponding numeric value; and (5) *overly restrictive definitions* for Mobility and Activity level, which referenced Braden Scale subscales and missed general functional assessments.

We broadened concept definitions where they were too restrictive, added concept-specific rules to disambiguate confusable concepts, tightened prompt guidelines for STRING types and null value handling, and added gating rules that require explicit tool or scale references for checklist items. The pipeline architecture, clustering, and post-processing remained unchanged.

The refined system improves F1 by 6.2 points on the development set (0.736→0.798) and 5.9 points on the test set (0.747→0.806). The gains are driven by precision (+12 on dev, +11 on test) with modest recall reduction (–2.4 on dev, –2.1 on test). NUMERIC precision improved from 0.742 to 0.929 after resolving the Oxygen saturation / Pulse oximetry confusion.

5.4. Error Analysis

Table 5 shows the top 5 false positive and false negative concepts in the initial system, alongside their counts after refinement.

Cat.	Concept	Initial	Refined
FP	Broset - confusion	102	0
	Patient identification	86	8
	Patient safety	81	35
	Delirium symptoms	80	18
	Pulse oximetry	67	4
FN	Mobility	23	14
	Dyspnea	18	20
	Fall risk ident.	18	21
	Delirium symptoms	17	17
	Temperature unit	15	15

Table 5: Top 5 false positive and false negative concepts from the initial system, with counts after refinement.

False Positives. The refinement reduced false positives across all five worst-offending concepts. Broset violence checklist confusion dropped from 102 to 0 after requiring explicit BVC references. Pulse oximetry dropped from 67 to 4 after disambiguating it from Oxygen saturation. Patient identification dropped from 86 to 8 after restricting extraction to specific identifiers. Patient safety and Delirium symptoms were also reduced, though they remain active sources of error.

False Negatives. False negative patterns were more stable across refinement. Mobility improved from 23 to 14 false negatives after broadening its definition beyond Braden Scale subscales. Dyspnea and Fall risk identification showed slight increases in false negatives (18→20 and 18→21). Delirium symptoms and Temperature unit were unchanged.

5.5. Discussion

The system extracts observations well from constrained value spaces. Refining definitions and rules from training-set error analysis further improves precision without changing the pipeline architecture. The largest gains came from resolving concept confusion between Oxygen saturation and Pulse oximetry, gating over-inferred behavioral concepts, and broadening restrictive definitions for functional assessments.

STRING extraction remains the bottleneck. The generative nature of LLMs produces semantically correct but lexically different text, which exact-match evaluation penalizes. Potential strategies include span selection from the transcript rather than free generation, fuzzy or semantic similarity matching in the evaluation, and few-shot prompting with gold annotation examples to calibrate output style.

5.6. Future Work

We plan to investigate: (a) confidence calibration to improve precision by filtering low-confidence predictions; (b) fine-tuning on the training data; (c) ensemble approaches across models, given headroom for model scaling; and (d) STRING-specific strategies such as span extraction.

6. Conclusion

We presented a modular LLM pipeline for the MEDIQA-SYNUR 2026 shared task that decomposes clinical observation extraction into four specialized stages. Medical domain clustering, knowledge-enhanced prompts, and deterministic validation yield strong results on structured value types. Training-set error analysis guided iterative refinement of definitions and rules, improving precision without sacrificing recall and achieving an F1 of 0.806 on the test set.

7. Bibliographical References

Asma Ben Abacha, Jean-Philippe Corbeil, et al. 2025. Overview of the MEDIQA-OE 2025 shared task on medical order extraction from doctor-patient consultations. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

Asma Ben Abacha, Jean-Philippe Corbeil, et al. 2026. Overview of the MEDIQA-SYNUR 2026 shared task on clinical observation extraction from nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop*. Forthcoming.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the BioNLP Workshop at ACL 2019*. Association for Computational Linguistics.

Asma Ben Abacha, Wen wai Yim, Yadan Fan, and Thomas Lin. 2023. Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

Asma Ben Abacha, Wen wai Yim, Meliha Yetisgen, and Fei Xia. 2024. Overview of the MEDIQA-CORR 2024 shared task on medical error detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 859–870, Suzhou, China. Association for Computational Linguistics.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2024a. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1949–1961.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Qianqian Xie, Yujia Zhou, Zhiyong Lu, Hua Xu, and Qiao Jin. 2024b. On-the-fly definition augmentation of LLMs for biomedical NER. *arXiv preprint arXiv:2404.00152*.

Juan C. Quiroz, Liliana Laranjo, Ahmet Baki Kocaballi, Shlomo Berkovsky, Dana Rezazadegan, and Enrico Coiera. 2019. Challenges of developing a digital scribe to reduce clinical documentation burden. *npj Digital Medicine*, 2(1):114.

Aaron A. Tierney, Gavin Gayre, Brian Hoberman, Bruce Mattern, Robert Muller, Marie Schmid, et al. 2024. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst*.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, 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*.

hgkai26 at MEDIQA-EVAL 2026: Automated Evaluation of Visual Medical Question Answering Using LLM-as-a-Judge

Haritha Gangavarapu
CGI Inc.
harithagangavarapu92@gmail.com

Abstract

As there is a rise in the use of multimodal large language models (LLMs) for medical response generation, it is necessary to have reliable automated evaluation mechanisms that can assess the quality of model-generated outputs. The MediQA-Eval 2026 shared task focuses on grading AI-generated dermatology and wound care responses using structured human-aligned rubrics. In this work, we explore a zero-shot multimodal LLM-as-a-Judge framework to assess candidate responses across multiple quality dimensions. System performance is evaluated using the official task metrics designed to reflect alignment with human judgments. Our findings provide preliminary insights into the feasibility and limitations of LLM-based evaluators for rubric-guided medical response assessment.
Keywords: LLM-as-a-Judge, Multimodal Evaluation, Dermatology and Wound Care QA

1. Introduction

Users often rely on online medical platforms to seek guidance on health-related issues. These platforms typically allow users to submit textual descriptions of their symptoms, and in certain specialties such as dermatology and wound care, accompanying clinical images are provided. Although medical professionals can respond to these questions, the high volume of submissions makes it challenging to provide timely responses.

To address scalability, AI systems can be developed to automatically generate responses based on user queries and associated images. However, automated responses in medical domains may contain inaccuracies, hallucinations, or unsafe recommendations. Since providing accurate information is the highest priority, before such responses can be presented to users, an additional evaluation layer is critical to assess their quality and safety.

The MediQA-Eval 2026 (mediQA26, 2026) evaluation task focuses on developing automated systems capable of grading AI-generated responses in dermatology and wound care settings. Each instance consists of a user query, associated images and multiple candidate responses generated by vision-language models. The objective is to build an evaluation system that assigns structured quality scores aligned with human judgments.

In this work, we implement a zero-shot multimodal LLM-as-a-Judge framework to automatically evaluate candidate responses. As part of the MediQA 2026 task, the hgkai26 team developed an automated evaluation system leveraging GPT-based models to grade three candidate responses from the dermatology and wound care datasets.

2. Related Work

2.1. Medical Multimodal Question Answering

Visual question answering datasets in dermatology and wound care were introduced in prior MEDIQA shared tasks, as part of MEDIQA-M3G (MediQA-M3G, 2024), MEDIQA-WV (MediQA-WV, 2025), and MEDIQA-MAGIC (ImageCLEF24, 2024) and 2025 (ImageCLEF25, 2025). These benchmarks consist of medical user queries along with associated images and have been used to study multimodal medical question answering.

2.2. Evaluation of Natural Language Generation Systems

Evaluating natural language generation (NLG) systems is inherently challenging, particularly in domains requiring domain expertise such as medicine. Traditional automatic metrics such as BLEU and ROUGE rely primarily on lexical overlap with reference texts. While these metrics are useful in structured generation tasks, they may not adequately capture semantic correctness, factual accuracy, or clinical safety in medical question answering(Celikyilmaz et al., 2021). In the evaluation of medical question answering systems, responses must be assessed not only for lexical overlap with reference texts, but also for factual accuracy and clinical safety. As a result, human evaluation using structured rubrics is often considered the gold standard. However, manual evaluation is expensive, time-consuming and difficult to scale. Many prior works have explored the wide use of large language models as automated evaluators (LLM-as-a-Judge) for NLG tasks (Liu et al., 2023; Xu, 2023; Saha et al., 2025).

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 257–261
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Split	Dataset	Queries	Systems	Responses	Metrics	Reviewers	Total Rows
Valid	iiyi	56	3	168	6	2	2016
Valid	woundcare	105	3	315	6	1	1890
Test	iiyi	100	3	300	6	2	3600
Test	woundcare	93	3	279	6	2	3348

Table 1: Number of instances by split for the English subset of the task dataset. The total number of rows corresponds to the product of the number of system responses, evaluation metrics, and reviewers.

3. Data

3.1. Task Dataset

The MediQA 2026 evaluation task builds upon existing dermatology and wound care visual question answering benchmarks. The dermatology subset consists of user queries originally posted on iiyi, a Chinese online medical consultation platform that allows users to ask their questions related to medical conditions or concerns and are answered by medical professionals. The wound care subset includes user queries and associated images collected from tieba.baidu.com and zhidao.baidu.com (wai Yim et al, 2025). Each instance includes three system-generated responses produced by Gemini-1.5-Pro, GPT-4o, and Qwen-VL-Chat using the original query and associated images as input context.

The dataset includes English (EN) and Chinese (ZH) subsets. The validation split contains 168 (iiyi) and 315 (wound care) English instances evaluated across six metrics, with corresponding Chinese subsets evaluated across two metrics, resulting in 3,864 human evaluation instances.

The test split contains 300 (iiyi) and 279 (wound care) English instances evaluated across six metrics, with corresponding Chinese subsets evaluated across two metrics, resulting in 4,632 human evaluation instances overall. Table 1 summarizes the number of annotation instances by split for the English subset.

In this work, we focus exclusively on the English evaluation track. To comply with task submission requirements, predictions for the Chinese (ZH) subset were assigned a placeholder value of -1 in our submission.

3.2. Annotator agreement

Given the subjective nature of response quality evaluation, we analyzed the inter-annotator agreement of the provided human annotations on the English test subset, where each instance was graded independently by two human reviewers. Table 2 reports agreement rates across evaluation metrics and datasets.

Inter-annotator agreement was computed locally by comparing the scores assigned by the two reviewers for each metric. A binary agreement score was assigned - 1 if both reviewers assigned the same rubric value, or 0 otherwise. Although the scoring rubric uses numeric values (e.g., {0, 0.5, 1}), these were treated as categorical labels for the purpose of agreement measurement.

As shown in Table 2, aggregate agreement differed substantially across datasets. The wound care subset exhibited higher *overall* agreement (81%) compared to the iiyi dermatology subset (56%).

For the iiyi dataset, writing style demonstrated comparatively lower agreement across systems. Metrics following the three-level rubric ({0, 0.5, 1}) showed lower agreement than the binary disagreement flag. In contrast, for the wound care dataset, writing style and relevance showed relatively higher agreement score. It is important to note that the reviewer groups for the iiyi and wound care datasets were not the same, which may partially contribute to the observed differences in agreement.

System	Metric	iiyi (n=100)	woundcare (n=93)
SYSTEM001			
	Completeness	0.48	0.87
	Disagree_flag	0.82	0.84
	Factual_accuracy	0.59	0.78
	Overall	0.57	0.74
	Relevance	0.66	0.99
	Writing_style	0.37	0.94
SYSTEM002			
	Completeness	0.48	0.87
	Disagree_flag	0.86	0.84
	Factual_accuracy	0.63	0.73
	Overall	0.66	0.69
	Relevance	0.69	0.98
	Writing_style	0.35	0.96
SYSTEM003			
	Completeness	0.30	0.62
	Disagree_flag	0.65	0.75
	Factual_accuracy	0.56	0.62
	Overall	0.55	0.59
	Relevance	0.61	0.85
	Writing_style	0.32	0.85

Table 2: Inter-annotator agreement across evaluation metrics and systems on the English test subset

4. Methods

4.1. Evaluation criteria

Each system response was evaluated across four dimensions: *completeness*, *factual accuracy*, *relevance*, and *writing style*. In addition, an *overall* score was assigned. Scores for these dimensions were given on a rubric of {0, 0.5, 1}, where 1 indicates correct, accurate, and fully relevant, 0.5 indicates a partially correct or incomplete response, and 0 indicates an incorrect or irrelevant response.

Additionally, a binary *disagree-flag* score was assigned, where 1 indicates disagreement with the candidate response and 0 indicates agreement.

These structured scores provide an overall assessment of the quality and reliability of the system response and help determine whether it is suitable to be presented to users.

4.2. GPT-as-a-Judge evaluation framework

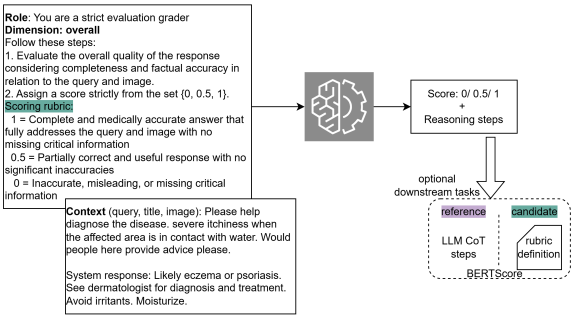


Figure 1: Example of a rubric-based instruction used for response evaluation.

We use a multimodal LLM-as-a-Judge setup based on OpenAI's `score_model` evaluation framework (OpenAIEvals, 2023). Each system response is graded across six metrics: *completeness, factual accuracy, relevance, writing style, overall, and disagreement flag*. For each instance, the grader receives the user query title (EN), query text (EN), associated image, and the candidate system response.

The model is instructed to act as a strict medical grader and assign a score from the set {0, 0.5, 1} according to rubric definitions provided in the prompt shown in Figure 1. As this is a zero-shot setup, no examples are provided to illustrate how the rubric levels 0, 0.5, and 1 should differ. The model relies solely on the metric specific rubric guidance when assigning scores.

We evaluated multiple GPT variants(OpenAI, 2024), including GPT-4o-mini and GPT-5-mini, under different parameter settings, including deterministic setting (temperature = 0). While our official submission included only the GPT-4o-mini run, we conducted additional experiments using GPT-5-mini after the test submission window closed. These additional evaluations were run locally using the official evaluation script.

In addition to numeric scores, the model also generates intermediate reasoning. We conducted preliminary analyses of these reasoning outputs to explore whether additional alignment signals could be extracted. However, as the observed performance was not sufficiently strong, these signals were not incorporated into the final system and are left for future work.

5. Results

We report results across the two datasets, iiyi and wound care, focusing on the *overall* metric as well as results computed on the combined data. To remain consistent with the task evaluation, we report the aggregated correlation metrics in Table 3. Additionally, in Table 4 we report classification accuracy for the *disagreement* dimension. Although accuracy was not part of the official evaluation metrics, we include it to quantify exact score agreement.

Kendall and Spearman measure rank-based correlation, while Pearson measures linear correlation. These metrics are appropriate for ordinal scoring settings, as they assess alignment between the system's score ordering and human judgments. In contrast, accuracy treats scores as discrete class labels (0, 0.5, 1) and measures exact agreement between system and human scores.

- ALL-en-overall-mean: The average of Kendall, Spearman, and Pearson correlations computed over all English data rows for the *overall* metric.
- ALL-en-ALL-mean: The average of Kendall, Spearman, and Pearson correlations computed over all English data rows and averaged across all evaluation metrics.

5.1. Observations on Zero-Shot Grading

While the overall correlation with human scores remained low across all configurations including stricter rubric prompts and the use of a larger model(GPT-4o), GPT-5-mini achieved higher correlation compared to the other setups. GPT-5-mini ranked weak system responses more similarly to the human reviewers. However, it assigned more conservative scores, particularly for the *overall* metric. In many cases where other models predicted a score of 1, GPT-5-mini predicted 0 or 0.5 instead. As a result, while classification accuracy remained low when treating scores as discrete labels, rank-based correlation metrics were relatively stronger.

Model	Params	Validation		Test	
		ALL-en-overall-mean	ALL-en-ALL-mean	ALL-en-overall-mean	ALL-en-ALL-mean
GPT-5-mini	default	0.3399	0.3095	0.3059	0.2547
GPT-4o-mini*	default	0.2811	0.2816	0.2444	0.2129
GPT-4o-mini	temperature=0	0.2845	0.3204	0.1757	0.2101

Table 3: Mean correlation results on the validation and test sets across all metrics and the Overall dimension. * indicates the official system submission

Model	Params	Disagree-flag Accuracy		Disagree-flag Mean Correlation	
		Woundcare	iiyi	Woundcare	iiyi
GPT-5-mini	default	**0.79**	**0.73**	**0.2251**	**0.2686**
GPT-4o-mini*	default	0.78	0.70	0.1866	0.2018
GPT-4o-mini	temperature=0	0.68	0.61	0.0902	0.1981

Table 4: Disagree-flag accuracy and mean correlation across datasets in the test split. * indicates the official system submission

We also observed that the deterministic setup (temperature = 0) achieved better correlation scores on the validation set. However, this improvement did not generalize to the test set, suggesting that deterministic setting alone does not consistently improve alignment with human evaluations. These findings suggest that the limitation lies not primarily in parameters or model scale, but in the zero-shot direct scoring setup itself. In this setting, the LLM judge must interpret the rubric definitions without any calibration examples. Which could be challenging to distinguish between fine-grained score levels, especially in medical evaluation tasks.

6. Conclusion

Our system explores a zero-shot rubric-based grading approach using the `score_model` evaluation type of the OpenAI evaluation framework. When the grading prompt included explicit structural constraints or deterministic settings, performance declined on the iiyi dataset, which also shows lower annotator agreement. Additionally, the accuracy for wound care with and without deterministic configuration (temperature = 0) was 0.78 vs 0.72, while for iiyi it was 0.49 vs 0.46, indicating lower agreement overall. Deterministic grading does not necessarily improve correlation with human judgments in multimodal medical evaluation. As a result, scoring occasionally exhibited variability across rubric boundaries. In future work, incorporating concrete examples to better distinguish rubric score levels and averaging across multiple runs may improve the stability of the evaluations. For the *overall* metric, GPT-5-mini provided more restrictive ratings compared to other setups, suggesting that restrictive scoring behavior may contribute to improved rank correlation.

7. Bibliographical References

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey.

ImageCLEF24. 2024. Imageclef 2024 medical: Mediqa. https://www.imageclef.org/2024/medical/mediqa. Accessed: 2026-02-18.

ImageCLEF25. 2025. Imageclef 2025 medical: Mediqa. https://www.imageclef.org/2025/medical/mediqa. Accessed: 2026-02-18.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

MediQA-M3G. 2024. Mediqa-m3g shared task. https://sites.google.com/view/mediqa2024/mediqa-m3g. Accessed: 2026-02-18.

MediQA-WV. 2025. Mediqa-wv shared task. https://sites.google.com/view/mediqa-2025/mediqa-wv. Accessed: 2026-02-18.

mediQA26. 2026. Mediqa-eval 2026. https://sites.google.com/view/mediqa2026/mediqa-eval. Accessed: 2026-02-18.

OpenAI. 2024. Gpt-4 technical report.

OpenAIEvals. 2023. Openai evals framework. https://github.com/openai/evals. Accessed: 2026-02-18.

Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. 2025. Learning to plan reason for evaluation with thinking-llm-as-a-judge.

Wen wai Yim et al. 2025. Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care. *Journal of Biomedical Informatics*, 170:104888.

Wenda et al. Xu. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Role-Adapted Clinical Report Generation for Ultrasound Measurements in Low-Resource Settings

Ayoub Nainia[1,2], **Tanya Akumu**[2], **Noussair Lazrak**[2], **Karim Lekadir**[2]
[1] Institut de Systématique, Évolution, Biodiversité (ISYEB), Sorbonne Université, Paris, France
[2] Artificial Intelligence in Medicine Lab (BCN-AIM), Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain
ayoub.nainia@sorbonne-universite.fr, tanya.akumu@ub.edu,
noussair.lazrak@ub.edu, karim.lekadir@ub.edu

Abstract

Obstetric ultrasound is critical for monitoring fetal growth, yet in many low-resource settings, healthcare workers who perform or receive ultrasound measurements lack the training to interpret them clinically. We present a system that automatically generates role-adapted clinical reports from fetal biometry measurements, targeting six healthcare worker roles across three expertise levels. The system combines Retrieval-Augmented Generation (RAG) from a knowledge base extracted from the World Health Organization (WHO) Manual of Diagnostic Ultrasound with deterministic fetal growth percentile computation based on INTERGROWTH-21st international standards. The knowledge base is designed for multilingual extensibility: since the source material is from an official WHO document, entries can be translated into any target language by domain experts or machine translation services. A key design principle is that clinical decision support (red, yellow, and green alerts) is derived deterministically from percentile thresholds, not from the language model, ensuring safety regardless of LLM output quality. Evaluation demonstrates sub-millimeter accuracy in percentile computation, 100% correctness in decision support classification, measurable readability differentiation across roles (Flesch-Kincaid grade 8.8 for community health workers vs. 11-13 for clinical roles), and 98% factual consistency across 42 generated reports spanning seven clinical scenarios. The system is designed for local deployment without internet connectivity.

Keywords: clinical NLP, low-resource settings, fetal ultrasound, report generation, decision support, RAG

1. Introduction

Fetal growth monitoring through ultrasound biometry is a cornerstone of antenatal care. Measurements such as head circumference (HC), biparietal diameter (BPD), abdominal circumference (AC), femur length (FL), and estimated fetal weight (EFW) are compared against population-based reference standards to identify growth abnormalities that may require clinical intervention (Papageorghiou et al., 2014).

However, in many low and middle-income countries (LMICs), the healthcare workers who obtain or receive these measurements, such as community health workers, nurses, and midwives, often lack the specialized training to interpret percentile values, identify abnormal growth patterns, or formulate appropriate clinical recommendations (World Health Organization, 2013). This interpretation gap means that even when ultrasound equipment is available, the clinical value of the measurements may be lost.

Large language models (LLMs) have shown promise in generating clinical text (Singhal et al., 2023), but deploying them in medical settings raises safety concerns: an LLM might generate plausible but incorrect clinical recommendations. This risk is amplified in low-resource settings where there may be no specialist available to catch errors.

We address these challenges with a system that:

1. Generates role-adapted clinical reports from ultrasound measurements, adjusting language complexity and clinical detail for six healthcare worker roles across three expertise levels (basic, intermediate, advanced).

2. Computes fetal growth percentiles deterministically using INTERGROWTH-21st equations (Papageorghiou et al., 2014; Stirnemann et al., 2020), ensuring that the quantitative assessment is independent of the LLM.

3. Provides deterministic decision support (red/yellow/green alerts) based on percentile thresholds, not LLM output, which is a critical safety design for contexts where the user cannot independently verify clinical recommendations.

4. Supports multilingual extensibility through a modular knowledge base extracted from the WHO Manual of Diagnostic Ultrasound, Volume 2. Since entries are sourced from an official WHO document, they can be translated into any language via professional translators or machine translation, enabling deployment across diverse linguistic settings.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 262–269
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Related Work

Large language models have advanced rapidly in medical applications, from clinical question answering (Singhal et al., 2023) to clinical text summarization, where adapted LLMs can match or outperform medical experts (Van Veen et al., 2024). Thirunavukarasu et al. (2023) review LLM deployment across clinical domains, identifying promising applications alongside persistent concerns about factual reliability. However, most work targets high-resource settings with specialist oversight, leaving a gap for environments where such oversight is unavailable.

For medical report generation specifically, earlier approaches used encoder-decoder architectures to produce text from medical images (Li et al., 2018; Chen et al., 2020). Retrieval-augmented generation (RAG) offers an alternative by grounding generation in authoritative sources (Lewis et al., 2020). Xiong et al. (2024) benchmarked medical RAG across five QA datasets, showing that retrieval quality critically determines whether RAG helps or harms performance. Our system uses RAG to ground narratives in WHO reference material, while relying on deterministic computation rather than the LLM for quantitative assessments.

A key challenge is that LLM-generated clinical text is prone to hallucination. Pal et al. (2023) introduced Med-HALT, the first benchmark for medical hallucinations, revealing significant variation across models. Jeblick et al. (2024) found that ChatGPT-simplified radiology reports occasionally contain incorrect or potentially harmful statements, even when the overall quality is high. Wiest et al. (2024) showed that locally-deployed open-source LLMs can achieve high accuracy for clinical information extraction while preserving data privacy, which is an approach our system also adopts. We go further by ensuring that safety-critical decision support is never derived from LLM output.

Despite these advances, most clinical NLP research remains focused on English-language EHRs (Electronic Health Records) in high-resource hospitals (Johnson et al., 2016). Ciecierski-Holmes et al. (2022) review AI deployments in LMICs and identify persistent barriers including the lack of representative datasets in under-resourced languages. For obstetric care specifically, Della Ripa et al. (2025) surveyed providers in African LMICs and found that clinical decision support alongside AI-enabled point-of-care ultrasound is deemed essential for improving antenatal outcomes, which is a need our system directly addresses.

Adapting medical text for different readers is well-established in health literacy research (Nutbeam, 2008). Recent work has applied LLMs to this task: Spina et al. (2024) demonstrated that GPT models can tailor information complexity to specific education levels, and Jeblick et al. (2024) showed ChatGPT can simplify radiology reports for patients. However, these studies focus on binary adaptation (expert vs. patient), whereas our system implements graduated adaptation across six healthcare worker roles at three expertise levels.

Our percentile computations build on the INTERGROWTH-21st international fetal growth standards (Papageorghiou et al., 2014) and the Hadlock formula for estimated fetal weight (EFW) (Hadlock et al., 1985), with EFW reference curves from Stirnemann et al. (2020).

3. System Architecture

Our system consists of four components: (1) a multilingually extensible knowledge base with retrieval pipeline, (2) an INTERGROWTH-21st percentile computation module, (3) a role-adapted prompt construction engine, and (4) a deterministic decision support layer. Figure 1 illustrates the overall architecture.

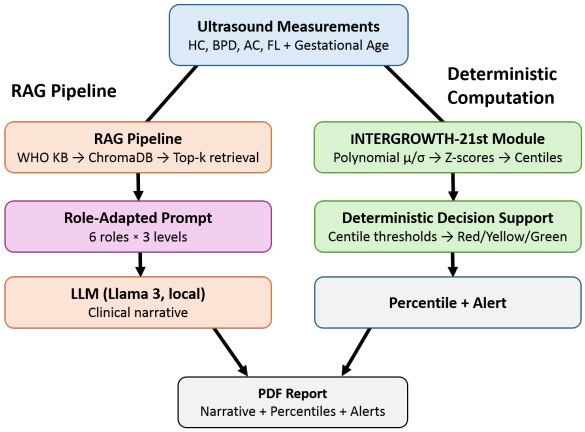


Figure 1: System architecture. Measurements flow through two independent paths: the RAG pipeline generates a clinical narrative via a locally-deployed LLM, while the INTERGROWTH-21st module computes percentiles that feed a deterministic decision support layer. The decision support alerts are never derived from LLM output.

3.1. Knowledge Base and RAG Pipeline

The knowledge base contains 171 entries extracted from the WHO Manual of Diagnostic Ultrasound, Volume 2 (World Health Organization, 2013). Each entry comprises a text passage with metadata (chapter, header, section, page number). Since the source material is from an authoritative WHO document, entries can be translated into any target language by professional translators for clinical-grade accuracy or via machine translation APIs for

rapid prototyping. As a proof of concept, we include machine-translated Swahili entries to demonstrate multilingual extensibility for East African deployment scenarios.

For retrieval, entries are encoded using the `paraphrase-multilingual-MiniLM-L12-v2` sentence transformer (Reimers and Gurevych, 2019) and stored in a ChromaDB vector database. Given a set of ultrasound measurements, the system generates search queries from measurement types and values, retrieves the top-k most relevant knowledge base entries, and includes them as context in the LLM prompt.

The architecture is LLM-agnostic: any model served through Ollama can be used as the generation backend. In our experiments, we use Llama 3 8B (Grattafiori et al., 2024) as it offers a good balance between capability and resource requirements for local deployment. For non-English output, a multilingual model variant can be used.

3.2. INTERGROWTH-21st Percentile Computation

Rather than relying on the LLM to interpret whether measurements are normal, we compute percentiles deterministically using the published INTERGROWTH-21st equations. For fetal biometry (HC, BPD, AC, FL, OFD), the z-score is computed as:

$$z = \frac{x - \mu(\mathrm{GA})}{\sigma(\mathrm{GA})} \tag{1}$$

where x is the measured value and μ and σ are gestational-age-specific polynomial functions derived from the INTERGROWTH-21st study. The centile is then obtained via the standard normal CDF: $C = \Phi(z)$.

For estimated fetal weight, we first compute EFW using the Hadlock formula from HC, AC, and FL measurements (Hadlock et al., 1985), then compute the z-score using the Box-Cox Power Exponential (BCPE) distribution parameters published by Stirnemann et al. (2020):

$$z_{\mathrm{EFW}} = \frac{(\mathrm{EFW}/\mu)^{\lambda} - 1}{\lambda \cdot \sigma} \tag{2}$$

where λ, μ, and σ are gestational-age-specific parameters. The valid gestational age range is 14^{+0} to 40^{+6} weeks for biometry and 18^{+0} to 40^{+6} weeks for EFW.

3.3. Role-Based Adaptation

The system defines six healthcare worker roles organized into three expertise levels, as shown in Table 1.

Level	Roles	Language Style
Basic	CHW	Plain language, no jargon, clear action items
Intermediate	Nurse, Midwife	Medical terms with explanations
Advanced	Clin. Officer, Sonographer, GP	Full terminology, differentials, mgmt. plan

Table 1: Healthcare worker roles and expertise levels. CHW = Community Health Worker, GP = General Practitioner.

Role adaptation is achieved through role-specific instructions injected into a structured prompt template. The prompt takes the form: *"You are a medical assistant helping a {role} analyze ultrasound measurements. Use ONLY the reference knowledge below."* It then provides (1) role-specific instructions, (2) the raw measurements with INTERGROWTH-21st percentile assessments, (3) retrieved knowledge base entries as reference context, and (4) general instructions requiring the LLM to use the percentile assessment as the primary reference, cite sources, and not invent information beyond the provided context.

Role-specific instructions control language complexity. For example, the Community Health Worker prompt instructs: *"Explain in simple, non-technical language [...] Use plain terms instead of medical jargon (e.g., 'baby head size' instead of 'head circumference'). End with a clear ACTION: either 'Continue routine care' or 'REFER to hospital'."* In contrast, the General Practitioner prompt requests *"full medical terminology, percentile interpretation, differential diagnosis considerations, and recommended management plan."* Intermediate roles (Nurse, Midwife) receive instructions that use standard medical terminology but include explanations of complex findings.

3.4. Deterministic Decision Support

A critical design decision is that clinical urgency classification is *never* derived from LLM output. Instead, a deterministic algorithm evaluates the INTERGROWTH-21st percentiles:

- **Red (URGENT):** Any measurement $<$ 3rd or $>$ 97th percentile.
- **Yellow (FOLLOW UP):** Any measurement in the 3rd-10th or 90th-97th percentile range (and none red).
- **Green (NORMAL):** All measurements between the 10th and 90th percentile.

The worst flag across all measurements determines the overall severity. Each role level receives a severity-appropriate message: basic-level workers receive simple referral instructions ("Refer to hospital immediately"), while advanced-level workers receive clinical guidance ("Consider detailed anatomical survey, Doppler assessment, and specialist referral").

This separation ensures that even if the LLM generates an inaccurate narrative, the action recommendation remains correct.

4. Evaluation

We evaluate six aspects of the system: (1) percentile computation accuracy, (2) decision support correctness, (3) report readability across roles, (4) factual consistency of generated reports, (5) qualitative role differentiation, and (6) comparison against a template baseline.

Evaluations 1-2 use the project's deterministic test suite: 94 unit tests that verify percentile computation against the published INTERGROWTH-21st reference tables and decision support classification against hand-labeled cases covering normal, borderline, and abnormal scenarios, including boundary conditions (e.g., a centile of exactly 3.0 classified as yellow vs. 2.9 as red). Of these, 17 specifically target the decision support classifier (Section 3.4).

Evaluations 3-6 use 42 reports generated by Llama 3 (8B, Q4_K_M quantization). We designed seven clinical scenarios that span the full severity spectrum: three normal cases at different gestational ages (16, 30, and 38 weeks) to test age-dependent interpretation, one borderline-low case (measurements near the 5th percentile), one abnormal-low case (measurements below the 3rd percentile), one case with multiple abnormal measurements, and one abnormal-high/macrosomia case (measurements above the 97th percentile). Each scenario was generated for all six roles, yielding $7 \times 6 = 42$ reports. While the dataset is modest in size, the scenarios were selected to cover every decision support category and clinical direction of abnormality (low and high), providing systematic coverage of the system's behavior space. Factual consistency was evaluated using automated negation-aware checks that verify agreement between each report's narrative content and the computed percentile data (details in Section 4.4).

4.1. Percentile Accuracy

We compared the equation-computed mean (μ) values against the official INTERGROWTH-21st percentile tables for gestational ages 14-40 weeks. Table 2 shows the results.

Measure	GA Range	Max Err. (mm)	Mean Err. (mm)
HC	14-40 wk	0.05	0.02
BPD	14-40 wk	0.05	0.02
FL	14-40 wk	0.05	0.03

Table 2: Equation accuracy vs. INTERGROWTH-21st reference tables (50th percentile) across $N = 27$ gestational age weeks per measurement.

Scenario	Cases	Correct	Acc.
Normal (10th-90th)	4	4	100%
Borderline (3-10/90-97)	5	5	100%
Abnormal (<3/>97)	4	4	100%
Red-overrides-yellow	1	1	100%
Boundary + edge cases	3	3	100%
Total	**17**	**17**	**100%**

Table 3: Decision support classification accuracy. All cases classified correctly, including boundaries (e.g., centile of exactly 3.0 $\rightarrow$ yellow; 2.9 $\rightarrow$ red).

HC, BPD, and FL show sub-millimeter accuracy across all 27 gestational age weeks, confirming faithful implementation of the published equations. The EFW centile computation was validated by round-trip testing: feeding reference table values (3rd, 50th, 97th percentiles) through the centile function yields maximum errors of 0.003 centile units for p50 across 19 GA weeks (22-40).

4.2. Decision Support Safety

We tested the decision support classification with 17 test cases covering normal, borderline, and abnormal scenarios, including edge cases at classification boundaries (Table 3).

All 17 cases were classified correctly, including the red-overrides-yellow propagation test.

4.3. Readability Across Roles

We computed standard readability metrics for all 42 generated reports. Table 4 shows the results per role, averaged across seven scenarios.

Reports for the basic-level role (CHW) are written at the lowest reading level (FK grade 8.8, corresponding to 8th-9th grade), while clinical roles (Nurse through GP) produce reports at FK grade 10-14 (college-level prose). The Flesch reading ease confirms this: CHW reports score 49.0, while most clinical roles score 27-37.

The Sonographer role is a notable exception among advanced roles, producing reports with a lower FK grade (10.2) and higher reading ease (42.6) than expected. Inspection reveals that Sonographer reports are structured as con-

Role	Level	FK Grade	FK Ease	Fog
CHW	basic	8.8	49.0	13.0
Nurse	interm.	11.2	36.5	15.4
Midwife	interm.	13.6	27.1	18.1
Clin. Officer	adv.	12.0	32.3	16.4
Sonographer	adv.	10.2	42.6	14.6
GP	adv.	11.9	31.3	16.7

Table 4: Readability metrics per role, averaged across 7 scenarios ($N = 42$ reports). FK = Flesch-Kincaid; FK grade and Gunning Fog correspond to US school grade levels; FK reading ease ranges from 0 (hardest) to 100 (easiest).

Scenario	Reports	Consistent	Acc.
Normal (16w/30w/38w)	18	18	100%
Borderline low	6	6	100%
Abnormal low	6	6	100%
Multiple abnormal	6	6	100%
Abnormal high (macrosomia)	6	5	83%
Total	**42**	**41**	**98%**

Table 5: Factual consistency of generated reports. Negation-aware matching distinguishes "no signs of growth restriction" from "growth restriction is suspected."

cise per-measurement technical assessments with short, declarative sentences (e.g., "The measured value of 279.0 mm falls within the normal range"), whereas GP and Clinical Officer reports use longer discursive paragraphs with subordinate clauses. This reflects a genuine difference in clinical communication style: sonographers typically produce structured technical findings, while clinicians produce narrative interpretations. The readability metrics thus capture both prompt-driven adaptation and role-appropriate discourse patterns.

4.4. Factual Consistency

We evaluated whether generated reports contradict the known percentile data using automated negation-aware checks across all 42 reports (Table 5).

Consistency checks verify that: (1) reports for normal measurements do not contain non-negated alarm terms (e.g., "urgent," "growth restriction"), (2) reports for abnormal measurements express appropriate concern, and (3) mentioned percentile values do not deviate from computed values by more than 30 percentage points.

Of 42 reports, 41 passed (98%). The single failure occurred in the macrosomia scenario (measurements $>$99th percentile): the Sonographer report mentioned "intrauterine growth restriction" despite all measurements being *above* the expected range. This error traces to the RAG pipeline, which retrieves growth restriction content for any abnormal measurement query without distinguishing the direction of abnormality. While the deterministic decision support correctly flagged this case as urgent, the LLM narrative contained clinically contradictory information, illustrating precisely why the system's safety-critical decisions are not derived from LLM output.

Community Health Worker (basic level):

"[...] The baby's head size is *smaller than normal* for this age. [...] I recommend **REFERRAL** to the hospital for further evaluation. This referral is needed because the baby's head size is significantly below expected."

General Practitioner (advanced level):

"[...] Head Circumference: 240.0 mm, significantly below the expected mean of 278.4 mm (0.0th percentile, z-score -4.15). [...] This may indicate a *small-for-gestational-age* (SGA) fetus. [...] Proper clinical management of intrauterine growth restriction requires maternal hospitalization and strict fetal surveillance, including a non-stress test and serial Doppler velocity waveform measurements."

Figure 2: Excerpted reports for the same abnormal scenario. The CHW report uses plain language and a clear REFER action; the GP report includes z-scores, differential diagnosis (SGA), and a detailed management plan.

4.5. Qualitative Role Differentiation

Figure 2 shows excerpted outputs for the same abnormal scenario (HC at 0th percentile, 30 weeks) generated for a Community Health Worker and a General Practitioner.

The CHW report uses accessible language ("baby's head size is smaller than normal") and ends with a clear action (REFER), while the GP report includes z-scores, references to SGA classification, and specific clinical management recommendations (Doppler assessment, non-stress test). This qualitative difference, combined with the quantitative readability results in Table 4, confirms that role adaptation is functioning as designed.

4.6. Template Baseline Comparison

To quantify the value added by the LLM over simple rule-based generation, we compared the RAG-generated reports against a template baseline that fills measurement values and decision support recommendations into fixed sentence patterns (Table 6).

LLM-generated reports are 2-4$\times$ longer than templates and contain nearly twice as many medical

Level	Source	FK Grade	Words	Med. Terms
Basic	LLM	8.8	196	5.3
	Tmpl	9.8	65	1.4
Interm.	LLM	12.4	251	9.2
	Tmpl	9.9	90	5.1
Adv.	LLM	11.3	333	9.0
	Tmpl	9.9	90	5.1

Table 6: LLM-generated reports vs. template baseline. Words = average word count; Med. Terms = count of 20 clinical terms (e.g., "differential," "Doppler," "SGA") found per report.

terms (5-9 vs. 1-5). Critically, the LLM produces meaningful readability differentiation between roles (FK grade 8.8 for basic vs. 12.4 for intermediate), while templates produce nearly uniform readability (FK 9.8-9.9) regardless of the target role. The templates also lack clinical context: they report measurements and a recommendation but cannot provide differential diagnosis considerations, management plans, or references to clinical guidelines that the RAG-grounded LLM incorporates.

5. Discussion

5.1. Design Principles for Clinical Safety

Our central design principle, that decision support must be deterministic, not LLM-generated, addresses a fundamental concern with deploying language models in clinical settings. The LLM provides valuable clinical narrative that contextualizes the measurements, but it does not determine whether a patient needs referral. This separation allows the system to be useful even when LLM output quality varies, which is particularly important when using smaller, locally-deployed models (8B parameters, 4-bit quantization) in low-resource settings.

The factual consistency evaluation (Section 4.4) provides evidence that this design works in practice: across 42 generated reports spanning seven scenarios, 41 were factually consistent (98%). The single failure, a Sonographer report that mentioned "intrauterine growth restriction" for a macrosomia case, illustrates a known RAG limitation: retrieval queries based on abnormal measurements do not distinguish the direction of abnormality. Importantly, this narrative error did not affect the deterministic decision support, which correctly classified the case as urgent. The template baseline comparison (Section 4.6) further validates the LLM's contribution: RAG-grounded reports provide 2-4$\times$ more content, meaningful readability differentiation, and clinical context (differential diagnoses, management plans) that static templates cannot offer.

5.2. Low-Resource Deployment

The system runs entirely locally using Ollama, requiring no internet connectivity or cloud API access. This is critical for deployment in settings with limited or unreliable internet infrastructure. The locally-deployed LLM also addresses data privacy concerns, as patient measurements never leave the local device.

The knowledge base architecture is designed for multilingual extensibility. Since entries are sourced from an official WHO document, they can be translated into any target language by local healthcare organizations or professional translators to produce clinically validated versions. The modular design allows new languages to be integrated without modifying the retrieval or generation pipeline.

5.3. Limitations

We have not yet conducted a user study with healthcare workers; readability metrics provide an objective proxy for language complexity but do not capture whether reports are understandable, actionable, or trustworthy for the intended audience. The factual consistency evaluation relies on negation-aware keyword matching, which covers the most critical error class (misidentifying normal vs. abnormal) but does not replace expert annotation. Although the knowledge base architecture supports multilingual deployment and we include machine-translated Swahili entries as a proof of concept, no direct evaluation of multilingual output quality was performed; validating retrieval and generation in non-English settings remains necessary before deployment. The knowledge base scope is also limited to 171 entries from a single WHO manual, which suffices for the fetal biometry use case but may lack coverage for more complex or ambiguous clinical scenarios. Finally, our evaluation uses a single model (Llama 3, 8B parameters); performance may vary across models. User studies, expert-annotated evaluation, multilingual evaluation, and multi-model comparison are planned as future work.

6. Conclusion

We presented a system for generating role-adapted clinical reports from obstetric ultrasound measurements, designed for healthcare workers in low-resource settings. The system combines RAG from a multilingually extensible WHO knowledge base with deterministic INTERGROWTH-21st percentile computation and a safety-focused decision support layer. Evaluation across 42 reports spanning seven clinical scenarios demonstrates: (1) sub-millimeter accuracy in percentile computation, (2) perfect decision support classification across 17 test cases,

(3) measurable readability differentiation across roles (FK grade 8.8 for community health workers vs. 11-14 for clinical roles), (4) 98% factual consistency with a single instructive failure that validates the deterministic safety layer, and (5) substantial improvement over template baselines in content richness and role differentiation. The system is designed for local deployment without internet connectivity.

Ethics Statement

This system is designed as a clinical decision support tool and is not intended to replace professional medical judgment. The deterministic decision support layer is designed to prioritize caution, recommending referral for any measurement outside normal ranges. The system has not been validated in a clinical setting and should not be used for clinical decision-making without appropriate validation and regulatory approval.

Acknowledgments

This research was funded by the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (AIMIX project – Grant Agreement No. 101044779). This work is co-funded by the European Union's Horizon Europe research and innovation programme through Cofund SOUND.AI under the Marie Skłodowska-Curie Grant Agreement No. 101081674.

7. Bibliographical References

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449. Association for Computational Linguistics.

Tomasz Ciecierski-Holmes, Raj Singh, Matthew Axt, Stefan Axt, and Samia Brenner. 2022. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. *npj Digital Medicine*, 5:162.

Sabrina Della Ripa, Christopher Oduor, Anthony Wanyoro, Peter Nguhiu, Elizabeth Rakers, and Cynthia K. Stanton. 2025. AI-enabled obstetric point-of-care ultrasound as an emerging technology in low- and middle-income countries: provider and health system perspectives. *BMC Pregnancy and Childbirth*, 25:729.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al. 2024. The Llama 3 herd of models.

Frank P. Hadlock, R.B. Harrist, Ralph S. Sharman, Russell L. Deter, and Seung K. Park. 1985. Estimation of fetal weight with the use of head, body, and femur measurements—a prospective study. *American Journal of Obstetrics and Gynecology*, 151(3):333–337.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, and Michael Ingrisch. 2024. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European Radiology*, 34:2817–2825.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

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. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in Neural Information Processing Systems*, volume 31.

Don Nutbeam. 2008. The evolving concept of health literacy. *Social Science & Medicine*, 67(12):2072–2078.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334. Association for Computational Linguistics.

Aris T Papageorghiou, Eric O Ohuma, Douglas G Altman, Tullia Todros, Leila Cheikh Ismail, Ann Lambert, Yasmin A Jaffer, Enrico

Bertino, Michael G Gravett, Manorama Purwar, et al. 2014. International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st project. *The Lancet*, 384(9946):869–879.

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. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, 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 Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Aidin Spina, Saman Andalib, Daniel Flores, Rishi Vermani, Faris F Halaseh, and Ariana M Nelson. 2024. Evaluation of generative language models in personalizing medical information: Instrument validation study. *JMIR AI*, 3:e54371.

Julien Stirnemann, Laurent J Salomon, and Aris T Papageorghiou. 2020. INTERGROWTH-21st standards for Hadlock's estimation of fetal weight. *Ultrasound in Obstetrics & Gynecology*, 56(6):946–948.

Arun James Thirunavukarasu, Daniel Shu Wei Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29:1930–1940.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30:1134–1142.

Isabella Catharina Wiest, Dyke Ferber, Jiefu Zhu, Marko van Treeck, Sonja K. Meyer, Radhika Juglan, Zunamys I. Carrero, Daniel Paech, Jens Kleesiek, Matthias P. Ebert, Daniel Truhn, and Jakob Nikolas Kather. 2024. Privacy-preserving large language models for structured medical information retrieval. *npj Digital Medicine*, 7:257.

World Health Organization. 2013. *Manual of Diagnostic Ultrasound*, 2nd edition, volume 2. World Health Organization, Geneva.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251. Association for Computational Linguistics.

A Comparative Study of Approaches to Anonymization of Clinical Free Text in Spanish

Florencia Brunello[1,2], **Serena Villata**[2], **Laura Alonso Alemany**[1], **Milagro Teruel**[1]
[1]Universidad Nacional de Córdoba, [2]Université Côte d'Azur
{florenciabrunello}@unc.edu.ar

Abstract

The anonymization of clinical free-text records is a prerequisite for enabling the secondary use of healthcare data while preserving patient privacy. This challenge is particularly acute for Spanish clinical text, where annotated resources are scarce and practitioners lack clear empirical guidance on which technological approaches are more adequate to their particular restrictions and capabilities.

In this work, we present a controlled comparative study of representative anonymization paradigms for Spanish clinical narratives, including a baseline rule-based approach, a general-purpose large language model under prompt-based inference, an off-the-shelf industrial NLP toolkit (spaCy) and comparable neural sequence labeling architectures. To ensure a fair and contamination-aware evaluation, particularly given the opacity of pretrained model training data, we introduce a synthetic clinical dataset designed to mirror the structural properties of established benchmarks while avoiding direct overlap.

Across heterogeneous evaluation settings, recurrent neural network architectures, particularly the off-the-shelf spaCy toolkit, consistently achieve the best balance between effectiveness, computational efficiency, and deployment feasibility. In contrast, the evaluated large language model and rule-based approach exhibit limited robustness in this domain. We further observe that training task-specific embeddings end-to-end yields stronger generalization than incorporating pretrained representations.

Although limited to Spanish and to representative instances of each paradigm, the study identifies stable performance tendencies across datasets. These results provide actionable guidance for institutions seeking to implement anonymization pipelines under realistic infrastructure constraints, and substantially narrow the technological search space for practitioners operating in under-resourced clinical NLP settings.

This work contributes reproducible evaluation procedures, synthetic benchmark resources, and empirical evidence for privacy-preserving clinical NLP in Spanish—an area where comparative studies remain limited.

Keywords: Clinical Text Anonymization, Spanish Clinical NLP, De-identification Methods Comparison

1. Introduction and Motivation

Clinical narratives such as discharge summaries, progress notes, and radiology reports contain rich, fine-grained information that is essential for clinical research, quality assessment, and healthcare optimization (Stubbs et al., 2015). However, these documents frequently include personally identifiable information (PII), making their use subject to strict legal, ethical, and regulatory constraints, including frameworks such as GDPR (European Parliament and Council of the European Union) and HIPAA (U.S. Congress, 1996).

De-identification, the removal or transformation of personally identifiable information to prevent re-identification of patients and workers, constitutes a core component of clinical data anonymization. It enables controlled data sharing across institutions and research initiatives while mitigating privacy risks. Although anonymization cannot eliminate all re-identification risks, it plays a central role in enabling the secondary use of healthcare data and supporting reproducible clinical research.

Despite its importance, anonymization remains a technically demanding task. Even manual de-identification has been shown to be time-consuming, cognitively demanding, and subject to inter-annotator variability (Dorr et al., 2006). Clinical text is characterized by high variability across institutions and specialties, frequent spelling errors, unconventional abbreviations, domain-specific shorthand, and elliptical phrasing. These properties make automatic anonymization particularly challenging and sensitive to domain shifts.

Over the years, a wide range of approaches has been proposed, including rule-based systems, statistical sequence models, and neural architectures. More recently, large pretrained language models (LLMs) have been applied to clinical de-identification, promising improved generalization and reduced need for feature engineering (Staab et al., 2024; Liu et al., 2025). At the same time, industrial NLP toolkits such as spaCy offer readily deployable pipelines that lower the entry barrier for institutions seeking practical solutions.

However, practitioners, particularly in under-resourced language settings, often face a fundamental question: which paradigm should be prioritized under realistic infrastructure and regulatory constraints? While English clinical NLP bene-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 270–280
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

fits from multiple comparative evaluations, Spanish clinical text remains comparatively under-studied. Annotated corpora are scarce, cross-dataset analyses are limited, and empirical guidance for technology selection is lacking. As a result, institutions frequently must make high-stakes technological decisions with limited evidence.

Evaluation practices further complicate this landscape. Performance is commonly measured on a small number of public benchmarks, such as i2b2 (Stubbs et al., 2015) and MEDDOCAN (**?**). For pretrained LLMs in particular, the opacity of training data raises concerns about benchmark contamination and potential information leakage, which may lead to overly optimistic performance estimates and hinder fair cross-paradigm comparison.

Robust evaluation therefore requires not only multi-dataset validation but also methodological safeguards against data leakage, especially when assessing large pretrained models. Additionally, clinical documentation varies substantially across institutions, languages, and document types. Systems that perform well on a single benchmark may fail to generalize in real-world deployment settings. For practitioners, this variability underscores the need for evaluation protocols tailored to local usage scenarios.

In this paper, we present a controlled and contamination-aware comparison of representative anonymization paradigms for Spanish clinical text. Specifically, we evaluate: (i) a baseline rule-based approach, (ii) general-purpose large language model under prompt-based inference, (iii) an off-the-shelf industrial NLP toolkit (spaCy), and (iv) fully trainable neural sequence labeling architectures based on recurrent neural networks. To mitigate benchmark contamination and ensure a fair comparison, we introduce two synthetic datasets that reproduce the structural and statistical characteristics of clinical narratives without replicating existing corpora. We complement these resources with evaluation on the recently released CARMEN-I dataset of naturally occurring Spanish clinical records (Farre Maduell et al., 2024).

By combining synthetic and natural datasets under uniform experimental conditions, this study aims to identify stable performance tendencies across paradigms rather than isolated benchmark scores. Our goal is not to claim universal superiority of any single architecture, but to provide empirically grounded guidance that narrows the technological search space for institutions implementing anonymization pipelines in Spanish clinical settings.

The remainder of this paper is structured as follows. Section 2 reviews related work in clinical de-identification, with special attention to Spanish-language resources. Section 3 describes the datasets and the rationale behind the synthetic data generation. Section 4 outlines the evaluation methodology, and Section 5 presents the anonymization approaches compared. Section 6 reports the experimental results. Finally, we discuss implications, limitations, and practical recommendations for deployment.

2. Relevant Work

Early work on clinical text anonymization relied primarily on rule-based systems, often implemented using regular expressions and domain-specific dictionaries (Neamatullah et al., 2008; Uzuner et al., 2007). These approaches offer transparency, deterministic behavior, and low computational cost, but they typically suffer from limited coverage, brittle pattern matching, and poor generalization to unseen writing styles or institutions because they are typically based on the concrete cases seen by the people who develop the systems, thus tailored to specific practices, and also because it seems impossible to have a rule-based system cover all use cases, as the resulting set of rules would most probably be inconsistent, with contradictory rules (Meystre et al., 2010). Similar tendencies have been observed in Spanish clinical text, where handcrafted rules are effective for highly unambiguous, regular entities such as dates or identification numbers, but struggle with ambiguity (specially abbreviations), lexical variability and informal writing (González et al., 2019).

Machine learning approaches, especially sequence-aware approaches like Conditional Random Fields (CRFs) and recurrent neural networks, later became dominant due to their ability to model contextual dependencies and reduce manual rule engineering, and, with deep learning, also reducing feature engineering (Dernoncourt et al., 2017; Lavergne et al., 2010). In the context of Spanish, shared tasks such as MEDDOCAN fostered the development of supervised sequence labeling systems for clinical de-identification (Martínez et al., 2019). Many top-performing systems in MEDDOCAN relied on bidirectional LSTM architectures, often combined with a CRF decoding layer, confirming the effectiveness of BiLSTM-CRF models for Spanish clinical named entity recognition (Saluja et al., 2019). These architectures build on earlier work demonstrating the effectiveness of deep-learning-based sequence labeling for NER (Lample et al., 2016; Ma and Hovy, 2016), and have since become a strong baseline or state-of-the-art for clinical anonymization tasks.

More recently, pretrained language models based on the Transformer architecture have been applied to de-identification, including for Spanish

medical text. Multilingual models such as mBERT and XLM-R, as well as Spanish-specific models like BETO, have shown promising results on clinical and biomedical NLP tasks (Cañete et al., 2020; Conneau et al., 2020; Moreno-Barea et al., 2025; Subies et al., 2025).

Within the MEDDOCAN challenge, several systems reported improved performance through the incorporation of contextualized embeddings and Transformer-based architectures (Martínez et al., 2019), highlighting the value of pretrained components in low-resource clinical settings. More recently, approaches leveraging large pretrained language models (LLMs) have been showing promising performance (Staab et al., 2024; Liu et al., 2025). In particular, prompt-based approaches have emerged as an attractive alternative, offering minimal implementation effort and rapid deployment, an appealing option for clinical institutions without specialized machine learning expertise.

However, these approaches present practical and methodological challenges. LLM-based systems entail substantially higher computational costs, which may be difficult to accommodate in secure, on-premise clinical infrastructures. In addition, because pretrained models rely on large and partially opaque training corpora, it is not always possible to determine whether reported benchmark performance reflects robust generalization or incidental exposure to similar data during pretraining (Bommasani et al., 2021). Consequently, performance reported on widely used shared-task datasets may not directly translate to new institutional contexts.

More broadly, anonymization systems frequently experience performance degradation when applied across institutions, document types, or stylistic conventions (Stubbs et al., 2017). This issue is particularly pronounced in Spanish, where publicly available resources remain scarce and often originate from a single institutional source (González et al., 2019).

These limitations motivate the present work, which conducts a controlled comparative evaluation across heterogeneous datasets to provide more reliable and practitioner-relevant evidence for Spanish clinical text anonymization.

3. Datasets

The main purpose in determining the datasets to be used in this evaluation was to ensure fair evaluation of the different approaches, trying to avoid data leakage from public datasets to large pretrained models, and to provide a low-effort methodology to create datasets that can be tailored to the specific characteristics of particular deployment contexts.

Four datasets were used for this comparison of anonymization approaches: the publicly available MEDDOCAN corpus, which has been available since 2019, and thus most probably used for pre-training of large language models, the 2025 CARMEN-I corpus, which has most probably not been used for training of models, and additionally presents a big variety of texts, and finally two synthetic datasets.

Sensitive entities are annotated following the MEDDOCAN Protected Health Information (PHI) guidelines (**?**), using a set of 28 entity types covering personal identifiers, healthcare professionals, institutions, locations, dates, contact information, identifiers, and other sensitive attributes.

3.1. MEDDOCAN

The MEDDOCAN dataset was released in the context of the MEDDOCAN anonymization of medical texts shared task organized within IberLEF 2019 (**?**). It consists of 1,000 synthetic clinical case reports manually selected and curated by medical experts and health documentalists, and enriched with realistic instances of protected health information derived from discharge summaries and medical genetics records (**?**).

3.2. Carmen-I

The CARMEN-I corpus consists of anonymized clinical narratives derived from electronic health records collected at Hospital Clínic de Barcelona between March 2020 and March 2022(**??**). For this study, we worked only with the 1,697 documents in Spanish. It comprises radiology reports (IR, 961 documents), discharge summaries (IA, 573 documents), transfer reports (IT, 154 documents), exitus reports (IE, 5 documents), and clinical course notes (CC, 4 documents).

3.3. Synthetic datasets

We created two synthetic datasets that reproduce the properties of the MEDDOCAN corpus, to better assess the performance of pre-trained components by ruling out data leakage, that is, that the test data had been used for training, in a manner similar to Kim et al. (2024).

The SPG dataset was generated using the Synthetic Patient Generator (SPG), an open-source software module developed for this work and released under the MIT license[1]. The generator produces fully synthetic clinical documents in Spanish, including protected health information such as names, addresses, dates, phone numbers, and identifiers, reproducing the MEDDOCAN style.

[1] repository URL withheld for double-blind review

Anonymization annotations were produced automatically by the generator in BRAT and XML formats.

Generation of reports was based on random sampling functions from predefined lexical resources containing person names, institution names, professions, and contact information. While this approach ensures full control over the generated content, it results in limited linguistic variability, as it relies exclusively on random selection from static lists and does not incorporate paraphrasing or syntactic rewriting mechanisms.

To increase linguistic variability and better approximate real clinical narratives, the SPG dataset was extended using the Llama 3.1 large language model. Starting from the original SPG documents, clinical reports were expanded to include richer medical details and longer narratives, while preserving the original patient and administrative sections.

The generation process followed a carefully designed one-shot prompting strategy enforcing a strict document structure and prohibiting meta-comments or safety disclaimers. Despite these constraints, the raw outputs required extensive manual curation. Documents containing refusals, meta-text, missing sections, or invalid structure were removed or corrected. After this curation process, a total of 448 extended clinical reports were retained.

4. Evaluation Methods

4.1. Dataset Partitioning

Each dataset was divided into three subsets: training (*train*), development (*dev*), and evaluation (*test*). For MEDDOCAN, CARMEN-I and SPG, a 50/25/25 split was used, following common practice in de-identification benchmarks. Due to the smaller size of SPG Extended after curation, an 80/10/10 split was applied. Table 1 summarizes the partition sizes and proportions for each dataset.

Dataset	Train	Dev	Test
MEDDOCAN	500 (50%)	250 (25%)	250 (25%)
CARMEN-I	847 (50%)	425 (25%)	425 (25%)
SPG	500 (50%)	250 (25%)	250 (25%)
SPG Extended	358 (80%)	45 (10%)	45 (10%)

Table 1: Train, development, and test splits for each dataset, in number of documents.

For CARMEN-I, the split was performed using stratified sampling by document type to preserve the original distribution of report categories (e.g., radiology, discharge summaries) across partitions. This stratification reduces bias toward dominant document types and supports better generalization.

4.2. Evaluation Metrics

All approaches were evaluated against gold standard annotations at the character level. Accuracy and F1-scores were obtained by macro-averaging the score for each class, leaving out the class of non-entities (the "O" class), which is a majority and tends to skew results.

Confusion matrices were also analyzed, to obtain a more fine-grained analysis of systematic prediction errors.

5. Approaches to anonymization

5.1. Rule-Based Anonymization Using Domain Knowledge

As a representative symbolic approach, we evaluated the set of regular expressions developed within the ARPHAI initiative[2]. These rules encode domain knowledge for detecting sensitive information in Spanish clinical texts and do not require any training data. They were designed to process free-text records of primary care in the public health system of Argentinean province La Rioja.

The system comprises two types of patterns: (i) context-aware expressions that detect entities directly from textual cues and their surrounding context, such as dates, national identification numbers, passport numbers, email addresses, and license numbers; and (ii) dictionary-based patterns that rely on curated lists of known entities, such as street names and hospital names.

5.2. Recurrent Neural Networks: BiLSTM-CRF

As a classical machine learning approach, we adopt the BiLSTM-CRF model implemented by Saluja et al. Saluja et al. (2019) for the MEDDOCAN shared task. The architecture follows the end-to-end sequence labeling approach proposed by Ma and Hovy (2016), composed of three main layers: CNNs, BiLSTMs, and a CRF. First, a Convolutional Neural Network (CNN) is used to extract character-level representations from words, allowing the model to capture morphological information like prefixes and suffixes. These character vectors are concatenated with pre-trained word embeddings and fed into a Bi-directional Long Short-Term Memory (BiLSTM) network, which models contextual information by processing the sequence in both forward and backward directions. Finally, the output vectors from the BiLSTM are passed to a Conditional Random Field (CRF) layer on top to jointly decode the optimal label sequence.

[2]https://www.ciecti.org.ar/arphai/

We hypothesize that, in the clinical domain and specifically for EHRs, the word representation component plays a critical role. Medical notes are frequently filled with typographical errors, non-standard abbreviations, and variations in word forms that rule-based systems often struggle to capture. By extracting morphological information (such as prefixes and suffixes) rather than looking for exact whole-word matches, the CNN layer allows the model to identify medical terms even when they are misspelled or appear as variations not seen during training. This design alleviates vocabulary sparsity and improves robustness to the noisy and informal nature of clinical text, compared to models that depend solely on word-level embeddings.

The model is trained in a fully supervised manner with the training partition of each dataset, using the validation partition to select the best hyperparameters.

The BiLSTM and CRF layers in our sequence labeling model are initialized with random weights; however, we also experiment with incorporating pretrained word embeddings. The hypothesis guiding this comparison is that embeddings learned from very large, general-purpose corpora encode linguistic and semantic regularities that may not be captured when training solely on smaller annotated datasets, even if they are domain-specific. We used the following:

- **PlanTL** word embeddings pre-trained for Spanish biomedical text with the Scielo corpus (Soares et al., 2019)
- **Flair** word embeddings for Spanish NER for the Flair toolset (Akbik et al., 2018)
- **spaCy** word embeddings for Spanish NER for the spaCy toolset (Honnibal et al., 2020)

5.3. Off-the-shelf Named Entity Recognition

In addition to custom models, we evaluated a widely used Named Entity Recognition (NER) tool, spaCy (Honnibal et al., 2020). It is a production-oriented natural language processing library that provides neural network-based models for text annotation. Its NER component is also based in a BiLSTM architecture, but it builds its representations using features derived from token vectors produced by the tok2vec layer and learned sequence patterns.

We initialized all the components of the spaCy model randomly using the default configuration provided by the library, which resulted in a training of the parameters *from scratch*.

Using spaCy offers several practical advantages. First, it is an off-the-shelf framework with a well-tested architecture similar to the previous one, which simplifies implementation and experimentation. Second, the tok2vec layer generates contextual token representations via a trainable embedding and more complex encoding subnetworks, potentially yielding richer features. However, the experiment is justified as further evaluation is necessary, since neural architectures trained from scratch may not generalize as effectively to new domains without sufficient labeled data.

5.4. Large Language Models

We evaluated a large language model approach using Meta-Llama-3.1-8B-Instruct, a transformer-based model optimized for instruction following. Experiments were conducted on GPU infrastructure at the *Centro de Computación Avanzada of the Universidad Nacional de Córdoba*, using mixed-precision inference to reduce computational costs.

Entity recognition was performed via prompt-based annotation, instructing the model to generate in-line XML annotations directly within the clinical text, as shown in Figure 1. Multiple prompting strategies were explored, including zero-shot and one-shot variants. Zero-shot prompting yielded inconsistent and poorly formatted outputs, whereas one-shot prompts with carefully curated examples significantly improved annotation quality.

Given the model's context window limitation (8192 tokens, including input and output), the prompt and input text were dynamically adjusted using the official tokenizer to ensure complete generation. Approximately half of the available context was reserved for the output to prevent truncation.

The Llama-based approach does not require task-specific training or fine-tuning, relying instead on prompt engineering to adapt the pretrained model to the anonymization task.

6. Analysis of Results

Table 2 shows macro-averaged accuracy and F1 scores computed on the test partition of each dataset. As discussed, all metrics exclude the O (outside) class in order to better highlight differences in the detection and classification of named entities.

It should be noted that, during the spaCy experiments, a small proportion of entities could not be processed due to formatting inconsistencies in the data. As a result, the reported scores are not strictly comparable. In most cases, the number of affected entities was below 5% and distributed among different labels, except in the case of the MEDDOCAN dataset and for label S̈EXO_SUJETO_ASISTENCIA, where it reached almost 10%. This issue led to a noticeable decrease in macro-averaged accuracy for that dataset,

```
<document>
  <sentence id="1">
    <TAG TYPE="NOMBRE_SUJETO_ASISTENCIA">Juan Perez</TAG>
    fue atendido el
    <TAG TYPE="FECHAS">12/03/2023</TAG>
    remitido por <TAG TYPE="NOMBRE_PERSONAL_SANITARIO">Ignacio</TAG>
    <TAG TYPE="NOMBRE_PERSONAL_SANITARIO">Navarro Cuellar</TAG>
  </sentence>
</document>
```

Figure 1: Example of XML-formatted clinical text with annotated entities.

	SPG		Extended SPG		MEDDOCAN		CARMEN-I	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
REGEX	0,094	0,092	0,081	0,092	0,155	0,123	0,004	0,001
BiLSTM	**0,942**	**0,884**	0,382	0,591	**0,864**	0,694	**0,702**	0,535
BiLSTM-PlanTL	0,817	0,820	0,472	0,560	0,632	0,690	0,347	0,420
BiLSTM-Flair	0,900	0,781	0,363	0,507	0,847	0,679	0,665	0,335
BiLSTM-spaCy	0,877	0,775	0,360	0,503	0,782	0,623	0,619	0,225
spaCy	0,846*	**0,880***	**0,588**	**0,675**	0,693	**0,770**	0,561	**0,657**
Llama 3.1	0,710	0,712	0,425	**0,674**	0,393	0,455	0,091	0,098

Table 2: Macro average accuracy and F1 excluding the non-entity class, for different approaches to anonymization: regular expressions (REGEX); different flavors of recurrent neural networks: BiLSTM fully trained on the training portion of each corpus, including the embeddings layer, and with different pre-trained embeddings (BiLSTM-PlanTL, BiLSTM-Flair, BiLSTM-spaCy); off-the-shelf spaCy fully trained on the training portion of each corpus; and off-the-shelf Llama 3.1. with prompts. Evaluation on three synthetic datasets, the public MEDDOCAN and the newly synthesized SPG and Extended SPG, and the naturally occurring CARMEN-I. Results marked with * correspond to a subset of data

although the F1 score remained largely unaffected. Given the exploratory nature of this study, we consider the results sufficiently reliable to identify general performance trends and to provide practitioners with meaningful guidance when selecting an anonymization approach.

As a first approach to these results, Figure 3 shows the differences in our main metric of interest, macro F1. Learning-based approaches substantially outperform REGEX (first bar in each group) across all datasets. In fact, for the only natural dataset, with real life variability, CARMEN-I, REGEX performs very poorly, reflecting its limited coverage and inability to generalize beyond explicit patterns. This could be expected given that these rules have been based on the study of a specific dataset, but it is sobering to determine empirically the very little portability this kind of approach, and how it requires an important adaptation effort for each new domain.

It is also very interesting to see how a generic pretrained LLM (last bar per group) performs very poorly for the natural dataset. In this case, we have a very important remark to make: the only case where Llama 3.1 performed comparably to other approaches was for Extended SPG, the synthetic dataset that had been created... using Llama 3.1. But such good performance in this dataset cannot be extrapolated to good performance on other datasets, or, more importantly, on a natural dataset like CARMEN-I. Thus, in working with synthetic datasets to assess performance of different technological alternatives, one must be very careful with the circularities of using an LLM to analyze data produced by the same or similar LLM.

The most important conclusion from this study is that **recurrent neural network architectures perform better for the task of automatic anonymization of clinical text**. The **best approach is off-the-shelf spaCy** (second-to-last bar per group), probably because

- all the components have been trained from scratch specifically for the dataset, and
- the architecture constructs a more complex representation of words and strings than classic BiLSTM, with a rich representation component transforming different kinds of embeddings to provide the input for the main recurrent component.

Indeed, it can be seen that training the embedding layers in the training dataset improves performance with respect to using pre-trained components, even if these have been trained in much

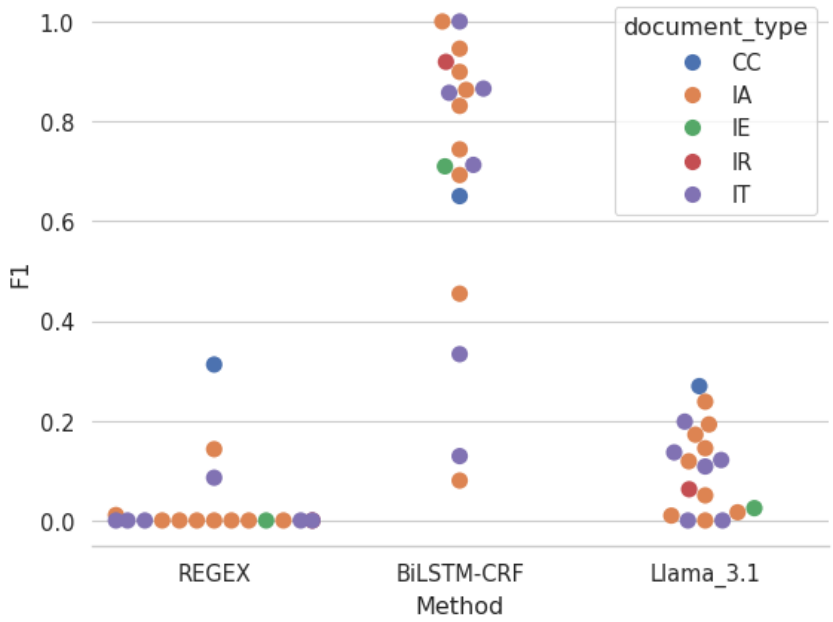


Figure 2: Distribution of performance of different anonymization approaches across different document types in the CARMEN-I datasets. Colors represent coarse types of documents, each dot represents a subtype of the coarser document types. The detail for each subtype of document will be provided in the camera-ready version of the paper, as an Appendix.

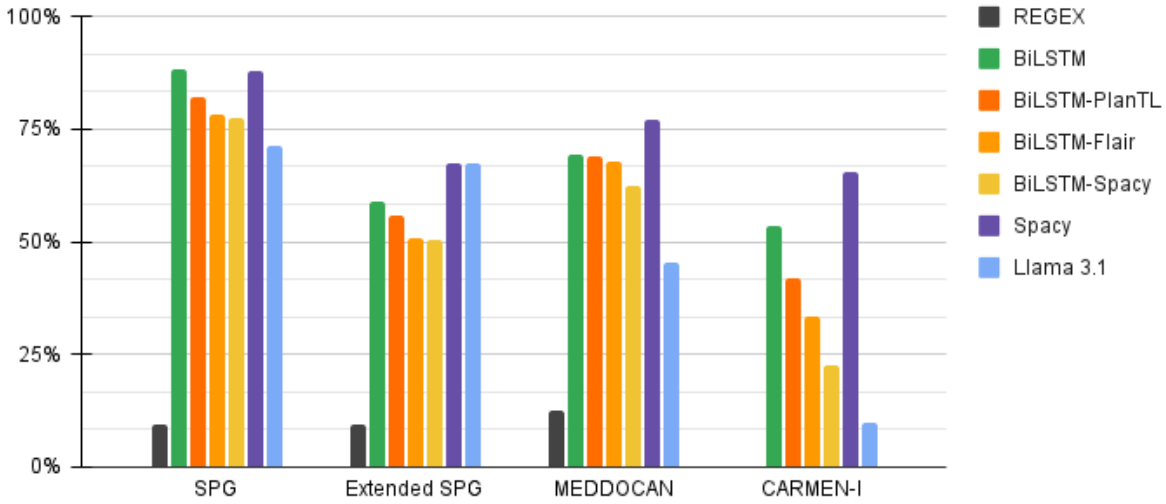


Figure 3: F1 macro of each approach on the different datasets, excluding the non-entity class. The detail can be seen in Table 2.

bigger corpora or a domain specific one. We can see that the BiLSTM that was fully trained on the datasets (second bar per group) performs consistently better than the BiLSTM approaches using pre-trained embeddings (3 middle bars per group). Even when using embeddings trained specifically in biomedical texts in Spanish, represented by the darkest orange bar never reach the performance of the fully trained BiLSTM. The only exception is the MEDDOCAN corpus, where performance of the different BiLSTM approaches is very similar. This can possibly be attributed to data leakage, since the MEDDOCAN corpus has been publicly available since 2019 and has been probably used for training these embeddings. More importantly, the drop in performance for approaches with pre-trained word embeddings is more acute for the CARMEN-I natural corpus, where the superiority of fully trained approaches can be clearly appreciated, followed by the relative better performance of the embeddings

pre-trained in biomedical domain.

Given the heterogeneity of the CARMEN-I corpus, we further analyzed performance by document type. Results in Figure 2 show that BiLSTM-CRF consistently outperforms REGEX and Llama 3.1 across most categories, although its performance is not uniform. As can be expected, the number of examples for a given document type strongly influences performance: indeed, the most notable degradation occurs for *IT_EXPLORACION_CLINICA*, a low-frequency document type with atypical date expressions, for which learned patterns do not generalize well.

7. Discussion and Conclusions

This study presents a controlled, cross-dataset comparison of representative anonymization paradigms for Spanish clinical text under uniform experimental conditions. Beyond reporting performance scores, our goal was to derive practically relevant insights for deployment in real clinical environments.

Across datasets, recurrent neural architectures, more concretely, CNN-BiLSTM-CRF, and in particular the spaCy implementation consistently outperformed both a rule-based system and a prompt-based approach with a large language model. The advantage was especially clear on heterogeneous, naturally occurring data such as that in the CARMEN-I dataset, which better reflects institutional variability.

From a deployment perspective, these models offer a favorable balance between effectiveness, computational cost, and infrastructural feasibility. They can be trained and executed on moderate hardware, including secure on-premise servers, whereas large language models require substantially greater computational resources and, if they are to be fine-tuned for a specific domain, skilled expertise. The prompt-based approach to a generic LLM, without fine-tuning, demonstrated limited robustness and weaker generalization to unseen natural data. Under realistic infrastructure constraints, moderately sized neural sequence labeling models therefore appear to be a more sustainable solution for Spanish clinical anonymization.

A consistent finding across datasets is that fully task-specific training of neural architectures yields stronger generalization than incorporating pretrained embeddings, even when those embeddings are derived from large biomedical corpora. This suggests that institutional specificity plays a central role: local documentation practices, abbreviations, and stylistic conventions are better captured through end-to-end training on representative annotated data. Importantly, effective anonymization does not require massive pretrained models. A well-curated, institution-specific dataset combined with a classical BiLSTM-CRF architecture can provide strong performance with lower operational complexity.

Methodologically, this work highlights the importance of contamination-aware and cross-dataset evaluation. The synthetic datasets introduced here were designed to replicate structural characteristics of clinical text without reproducing existing benchmarks, enabling more controlled comparisons. At the same time, results show that synthetic data must be handled carefully, as models may benefit when evaluated on data generated using related architectures. Complementary evaluation on naturally occurring corpora such as CARMEN-I is therefore essential. Performance differences between MEDDOCAN and CARMEN-I further demonstrate that single-benchmark evaluation is insufficient to guarantee real-world robustness, particularly in Spanish clinical NLP, where resources are scarce and institution-specific.

Taken together, these findings provide indicative but stable tendencies that can guide practitioners. While this study is limited to Spanish and to representative instances of each paradigm, it narrows the technological search space for institutions implementing anonymization pipelines under realistic constraints. In Spanish clinical contexts, investment in moderately sized, task-specific neural sequence labeling models appears to offer the strongest balance between effectiveness, efficiency, and deployability.

Beyond empirical comparisons, this work contributes reproducible evaluation procedures, synthetic benchmark resources, and cross-dataset evidence in an underrepresented language setting. By foregrounding methodological rigor and deployment feasibility, we aim to support more informed, evidence-based decisions in privacy-preserving clinical NLP.

Future work should extend this framework to additional languages and institutional contexts, explore domain-adapted or fine-tuned large language models, and investigate hybrid or lightweight architectures that further balance interpretability, robustness, and infrastructural constraints.

Ethics Statement

This work addresses the anonymization of clinical free-text records with the objective of enabling privacy-preserving secondary use of healthcare data. The study focuses on methodological evaluation and does not involve direct access to identifiable patient information by the authors.

Experiments were conducted using (i) publicly available benchmark datasets, (ii) the CARMEN-I corpus accessed under its Data Usage Agreement,

and (iii) synthetic datasets generated to replicate structural and statistical properties of clinical narratives without reproducing real patient content. The CARMEN-I resource consists of previously anonymized clinical documents distributed under controlled conditions via PhysioNet. All usage complied with the corresponding licensing and data access requirements.

The synthetic datasets were designed to mitigate potential benchmark contamination and reduce privacy risks associated with redistributing sensitive clinical material. Nevertheless, synthetic data cannot entirely eliminate the possibility of unintended memorization effects in large pretrained models. For this reason, we explicitly separate data generation and evaluation pipelines and complement synthetic evaluation with experiments on naturally occurring clinical data.

We acknowledge that anonymization systems are not infallible and may fail to detect certain categories of sensitive information. In real-world deployment, automated de-identification systems should therefore be integrated within broader governance frameworks that include human oversight, identification and charcterization of discriminatory behaviors including bias against subgroups of population, risk assessment, and compliance with applicable data protection regulations.

Finally, this work focuses exclusively on Spanish clinical text, a comparatively under-resourced setting in clinical NLP. By providing reproducible evaluation procedures and comparative evidence, we aim to support responsible technological decision-making in privacy-preserving healthcare applications (López et al., 2023).

Limitations

Several limitations must be acknowledged.

First, this study focuses exclusively on Spanish clinical text. Although this language-specific scope addresses the scarcity of annotated resources and comparative research in Spanish clinical anonymization, it limits the generalizability of our findings to other languages. Differences in documentation practices, annotation schemes, and linguistic features may affect performance, and further studies are needed to assess cross-linguistic robustness.

Second, although the synthetic datasets were carefully designed to resemble clinical narratives, synthetic data cannot fully replicate the linguistic variability, stylistic idiosyncrasies, and implicit conventions of naturally occurring documentation. Consequently, performance estimates obtained on synthetic corpora may differ from those observed in real-world deployment scenarios.

Third, the evaluation of large language models was restricted to a single representative LLM under prompt-based inference, without task-specific fine-tuning. Similarly, the off-the-shelf comparison was limited to spaCy as a representative general-purpose industrial NLP toolkit. While these choices reflect realistic scenarios faced by practitioners, who often evaluate a small number of accessible tools, they do not exhaust the full spectrum of possible architectures, fine-tuning strategies, or tool configurations. Alternative LLMs, domain-adapted transformers, or additional industrial frameworks may yield different absolute performance levels.

Finally, spaCy experiments were affected by preprocessing and formatting issues that prevented processing of a subset of documents. Although the overall performance trends remain stable across datasets and configurations, exact metric comparisons should therefore be interpreted with caution.

Acknowledgements

This study was partially supported by ARPHAI-CIECTI, a project funded by IDRC (Canada) and SIDA (Sweden). This work used computational resources from UNC Supercómputo (CCAD) – Universidad Nacional de Córdoba (https://supercomputo.unc.edu.ar), which are part of SNCAD, República Argentina.

Bibliographical References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1638–1649. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *Proceedings of PML4DC at ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, et al. 2020. Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL*.

Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*

(EMNLP), pages 199–204. Association for Computational Linguistics.

David A. Dorr, William F. Phillips, Shyam Phansalkar, Stephanie A. Sims, and John F. Hurdle. 2006. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods of Information in Medicine*, 45(3):246–252.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council.

Eulalia Farre Maduell, Salvador Lima-Lopez, Santiago Andres Frid, Artur Conesa, Elisa Asensio, Antonio Lopez-Rueda, Helena Arino, Elena Calvo, Maria Jesús Bertran, Maria Angeles Marcos, Montserrat Nofre Maiz, Laura Tañá Velasco, Antonia Marti, Ricardo Farreres, Xavier Pastor, Xavier Borrat Frigola, and Martin Krallinger. 2024. CARMEN-I: A resource of anonymized electronic health records in Spanish and Catalan for training and testing NLP tools. *PhysioNet*. Version 1.0.1.

Ana González, David Martínez, Sergio Montalvo, et al. 2019. Overview of the meddocan track: Medical document anonymization. *Proceedings of the IberLEF Workshop*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo*.

Woojin Kim, Sungeun Hahm, and Jaejin Lee. 2024. Generalizing clinical de-identification models by privacy-safe data augmentation using GPT-4. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21204–21218, Miami, Florida, USA. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *Proceedings of NAACL-HLT*, pages 260–270.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale crfs. *Proceedings of the 48th Annual Meeting of the ACL*, pages 504–513.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, Fang Zeng, Lichao Sun, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. 2025. Deid-gpt: Zero-shot medical text de-identification by gpt-4.

Sabrina López, Laura Alonso Alemany, Juan Manuel Dias, Laura Ación, and Verónica Xhardez. 2023. Guía práctica para la protección de datos personales en salud. Technical report, Fundar, Buenos Aires.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *Proceedings of ACL*, pages 1064–1074.

David Martínez, Ana González, and Sergio Montalvo. 2019. Meddocan: Medical document anonymization shared task. *IberLEF 2019 Working Notes*.

Stéphane M. Meystre, Jeff Friedlin, Brett R. South, Shengyu Shen, and Matthew H. Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, 10(1):70.

Francisco J. Moreno-Barea, Guillermo López-García, Héctor Mesa, Nuria Ribelles, Emilio Alba, José M. Jerez, and Francisco J. Veredas. 2025. Named entity recognition for de-identifying spanish electronic health records. *Computers in Biology and Medicine*, 185:109576.

I. Neamatullah, M. M. Douglass, L. H. Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark, and G. D. Clifford. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):32.

Binit Saluja, Ankit Kumar, Ankit Suri, and S Onur. 2019. Anonymization of clinical text in spanish using bilstm-crf. In *Proceedings of IberLEF*.

Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger, and Jordi Armengol-Estapé. 2019. Medical word embeddings for spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 124–133.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Large language models are advanced anonymizers.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. Practical applications for deep learning in clinical natural language processing. In *Proceedings of the BioNLP 2017 Workshop*, pages 25–34. Association for Computational Linguistics.

Amber Stubbs, Colton Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task

track 1. *Journal of Biomedical Informatics*, 58(Suppl):S11–S19.

Guillem García Subies, Álvaro Barbero Jiménez, and Paloma Martínez Fernández. 2025. ClinteXtsp and rigoberta clinical: a new set of open resources for spanish clinical nlp.

U.S. Congress. 1996. Health insurance portability and accountability act of 1996. Public Law 104-191, 110 Stat. 1936.

Ozlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2007. Identifying patient information in medical records. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Disagreement-Driven Joint Refinement of Retrieval and Decision Rules for Imbalanced Counseling Risk Classification

Zhihao Shao[1], **Ryo Sekizaki**[2], **Shengzhou Yi**[1], **Toshihiko Yamasaki**[1]
[1]The University of Tokyo
[2]Welcome to Talk Co., Ltd.
{shao,yishengzhou,yamasaki}@cvm.t.u-tokyo.ac.jp
sekizaki-ryo@welcometotalk.co.jp

Abstract

With the rapid growth of online counseling services, timely and reliable risk classification of counseling records is essential for supporting early screening and prioritizing limited intervention resources. High-risk samples refer to high-acuity suicide risk and require expedited human review. However, this task is challenging due to severe class imbalance (93% low-risk and 7% high-risk samples) and complex decision boundaries. Large language models (LLMs) exhibit unstable predictions and systematic errors in such imbalanced clinical-text settings. To address this issue, we propose Disagreement-Driven Joint Refinement (DDJR), an iterative, parameter-free refinement framework. It uses prediction disagreement between two inference settings, zero-shot and retrieval-augmented in-context learning, as the primary signal for identifying high-value instances. These disagreement-identified instances are transformed into adaptive refinement signals and used to jointly update both the exemplar pool and an executable rule set, thereby sharpening decision boundaries and improving prediction stability. Experiments on 6,481 real-world counseling records demonstrate that the proposed DDJR outperforms existing methods, achieving an accuracy of 0.915 and a Matthews Correlation Coefficient (MCC) of 0.583. These results demonstrate that DDJR achieves more stable and reliable predictions for high-stakes counseling risk classification in real-world settings.

Keywords: Mental Health Intervention, Large Language Models, Self-correction, Exemplar Retrieval

1. Introduction

The World Health Organization (WHO) estimates that approximately 970 million people worldwide are affected by mental disorders. Among these, depression and anxiety are the most prevalent, and suicide remains one of the leading causes of death among individuals aged 15 to 29 (World Health Organization, 2023; Shen et al., 2024). Text-based online counseling has improved access to mental health support while offering greater privacy. However, the absence of nonverbal cues in text-only interactions means that emotional assessment must rely primarily on linguistic signals (Park et al., 2019; Zhang et al., 2022). Therefore, reliable analysis of counseling dialogues is critical for triage and early intervention, as high-risk samples may require expedited human review. In this work, counseling risk is defined in terms of suicide-risk acuity: high-risk denotes high acuity, whereas low-risk denotes low acuity.

As illustrated in Figure 1, automatic counseling risk classification of dialogues can support professionals by identifying high-risk samples and prioritizing them for intervention. Early rule-driven and statistical methods, followed by neural models such as recurrent neural networks (Elman, 1990), improved linguistic modeling but remained limited in capturing long-range context and complex affective dependencies (Kermani et al., 2025). More recently, transformer-based models and large language models (LLMs) have strengthened contextual modeling, motivating increased interest in applying LLMs to counseling risk classification in mental health settings (Matero et al., 2019; Li et al., 2024). In this domain, interpretability and evidence grounding are critical. Nevertheless, counseling dialogue classification remains challenging because distress signals are often indirect and distributed across multiple turns, while LLMs may overemphasize surface cues or generate weakly supported rationales. This challenge is further exacerbated by severe class imbalance, as high-risk samples are rare but clinically critical.

To address these challenges, we propose Disagreement-Driven Joint Refinement (DDJR), an iterative inference-time optimization framework for imbalanced counseling risk classification. Our main contributions are as follows:

1. Disagreement-driven contrastive typing supervision is proposed to assign each instance to one of three types based on prediction disagreement, which provides signals for refinement without parameter updates.
2. Joint refinement of the exemplar pool and the rule set is proposed to enable closed-loop refinement without parameter updates.
3. Type-weighted exemplar retrieval is proposed to prioritize discrepancy and failure-oriented evidence from samples with different learning values, improving calibration for imbalanced counseling risk classification.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 281–289
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

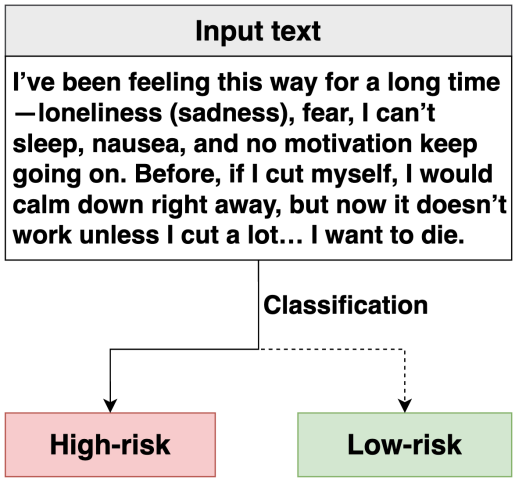


Figure 1: Suicide-risk acuity classification for counseling dialogues.

2. Related Work

2.1. Mental Health Intervention Systems

Dialogue-based mental health intervention systems have been explored as scalable support in text-based settings. Early systems such as Woebot (Fitzpatrick et al., 2017) demonstrated the feasibility of providing cognitive behavioral therapy-inspired guidance through scripted dialogue management and lightweight personalization. Subsequent reports on Tess (Rauws et al., 2019) described affect-aware interactions through emotion detection from user messages together with multilingual delivery. Systematic evidence indicated that mental-health chatbots could improve symptoms in some settings, but they also revealed substantial heterogeneity in study quality and limited safety reporting (Abd-Alrazaq et al., 2020). More broadly, conversational agents in healthcare remained an emerging area in which rigorous evaluation protocols and safety assessments have not yet been consistently standardized (Laranjo et al., 2018). Within psychiatry, recent reviews have further summarized the use of chatbot for both screening and intervention, emphasizing gaps in real-world validation and reporting (Vaidyam et al., 2019).

Recent work with large language models explored reasoning-oriented prompting for empathetic response generation, including Chain-of-Empathy (CoE), which performed explicit emotion analysis before response generation (Lee et al., 2023), and approaches that integrated rational cognition with value-oriented dialogue to promote well-being (Mercer, 2022). In contrast, our work targets risk triage for counseling dialogues, where decision stability under severe class imbalance and evidence alignment were central requirements for safe escalation to human review.

2.2. Retrieval-Augmented In-Context Learning for Counseling Dialogues

In high-stakes settings, it was desirable to ground model outputs in inspectable external sources rather than relying solely on parametric memory, thereby motivating retrieval-augmented approaches for more reliable and transparent predictions (Guu et al., 2020). Beyond open-domain retrieval-augmented generation (RAG) (Lewis et al., 2020), retrieval-augmented in-context learning (RA-ICL) retrieved labeled exemplars and uses them as in-context demonstrations.

Retrieval-augmented pipelines have been studied as a grounding strategy in mental health text classification, aiming to improve reliability and produce more explainable outputs (Kermani et al., 2025). However, RA-ICL was highly sensitive to demonstration selection. Dr.ICL, a retrieval-based in-context learning method, showed that retrieval-based demonstrations could outperform random sampling (Luo et al., 2024).

Motivated by this sensitivity, DDJR improves robustness under severe class imbalance by refining the RA-ICL exemplar pool to better capture discrepancy- and failure-oriented evidence, mitigating the under-coverage of hard cases during retrieval.

2.3. Self-Refinement in LLMs

Self-improvement strategies for LLMs can improve behavior either at inference time without parameter updates or through self-training with parameter updates. Reflection-based methods such as Reflexion (Shinn et al., 2023) and SELF-REFINE (Madaan et al., 2023) generated critiques to guide iterative revision without changing model weights, while STaR (Zelikman et al., 2022) improved reasoning by generating rationales and iteratively fine-tuning the model on them.

Recent work has further introduced structured self-verification to improve reliability. ProCo (Wu et al., 2024) used key-condition masking and iterative verification to refine outputs in complex reasoning tasks, and PAG (Jiang et al., 2025) alternated between policy and generative-verifier roles to detect errors and revise outputs across multi-turn interactions. Constraint-guided frameworks such as Constitutional AI externalized decision rules to produce more stable and interpretable outcomes (Bai et al., 2022).

In contrast, DDJR targeted risk classification in counseling dialogues under severe class imbalance. It uses disagreement signals to jointly refine the exemplar pool and the rule set, thereby improving evidence alignment for risk triage.

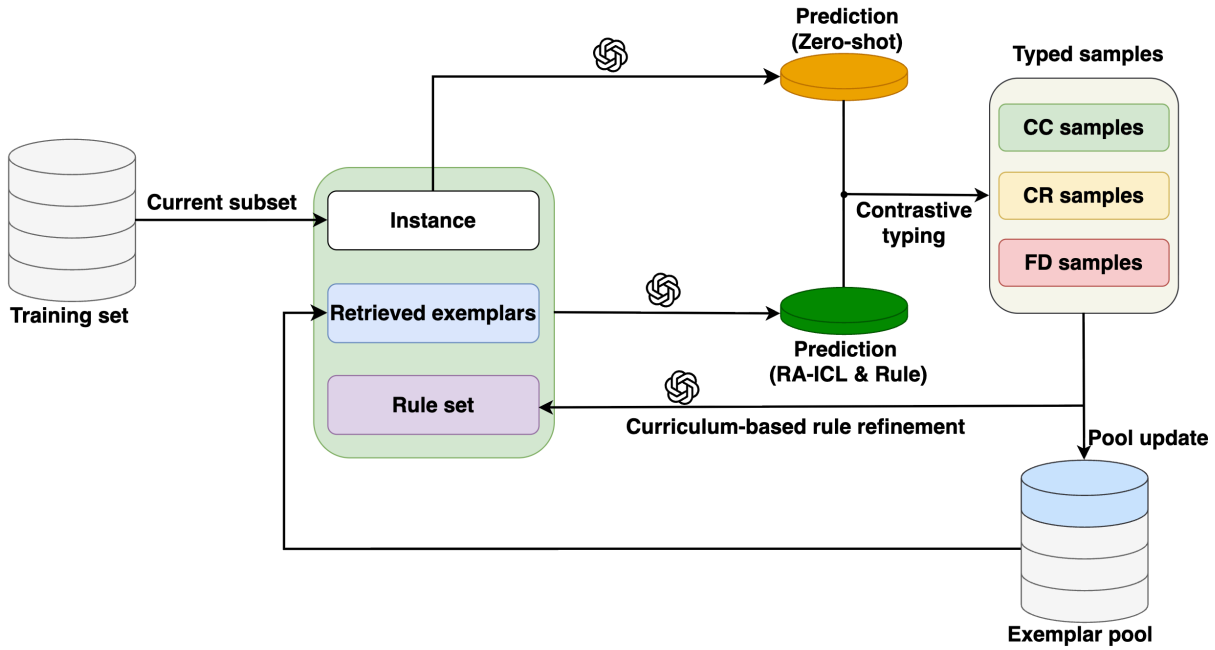


Figure 2: The overview of the Disagreement-Driven Joint Refinement (DDJR) framework.

3. Methodology

3.1. Overview

As shown in Figure 2, the proposed framework is a parameter-free iterative refinement loop that operates on the training set without model weight updates. It consists of three steps: typing instances, updating an exemplar pool, and refining a rule set. The exemplar pool and the rule set are updated jointly across refinement rounds.

3.2. Contrastive Typing Supervision

We obtain typed supervision signals by running dual inference for each labeled instance and using prediction disagreement in correctness outcomes as the primary refinement signal.

Let the training dataset be denoted as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{M} \tag{1}$$

where x_i represents a counseling dialogue and $y_i \in \{0, 1\}$ denotes its risk label, with 0 indicating low-risk and 1 indicating high-risk.

For each input text x_i, we run dual inference: (i) zero-shot inference, yielding prediction $\hat{y}_i^{(0)}$; and (ii) RA-ICL inference, yielding prediction $\hat{y}_i^{(1)}$. The goal is to derive structured supervision signals for exemplar pool and rule refinement, rather than performing prediction ensembling.

To characterize the outcomes of dual inference, we define a correctness indicator for each configuration:

$$c_i^{(k)} = \mathbf{1}\big(\hat{y}_i^{(k)} = y_i\big), \quad k \in \{0, 1\}, \tag{2}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. We then construct a correctness vector

$$c_i = \big(c_i^{(0)}, c_i^{(1)}\big) \in \{0, 1\}^2 \tag{3}$$

By construction, each instance must fall into exactly one of the following mutually exclusive categories:

$$c_i \in \{(1, 1)\} \cup \{(1, 0), (0, 1)\} \cup \{(0, 0)\} \tag{4}$$

Note that this contrastive typing requires gold labels and is performed offline on labeled data during the refinement stage.

Based on this correctness vector, we define three types of supervision signals:

(i) Consistently-Correct samples. When $c_i = (1, 1)$, both configurations predict correctly. We directly construct a standard Consistently-Correct (CC) sample (x_i, y_i). Such instances reflect consistent behavior across prompting conditions and typically correspond to relatively clear decision boundaries.

(ii) Contrastive Rationalization samples. When $c_i \in \{(1, 0), (0, 1)\}$, exactly one configuration predicts correctly. We construct a Contrastive Rationalization (CR) sample consisting of the input text, the ground-truth label, and two rationales: one generated by the correct configuration and the other by the incorrect configuration. Because the input is identical while correctness differs, these samples

highlight discrepancies in evidence selection and contextual interpretation, providing supervision on reasoning patterns to encourage or discourage.

(iii) Failure-Diagnostic samples. When $c_i = (0, 0)$, both configurations predict incorrectly. We treat the instance as a hard case and construct a Failure-Diagnostic (FD) sample. In this setting, we provide the gold label together with the incorrect rationales, enabling the model to summarize the likely missing cues, misleading cues, or boundary confusions shared by both failed predictions. Each FD instance therefore comprises the input text, the ground-truth label, and the generated error diagnosis.

Through this correctness-vector induced mechanism, CC samples capture stable patterns, CR samples provide contrastive reasoning supervision, and FD samples summarize shared failure modes. Together, they form a structured supervision pool that supports both rule refinement and exemplar pool updates.

3.3. Exemplar Pool Refinement

The few-shot exemplar pool is initialized with the full training set and its size is kept fixed across iterations. In each iteration, every newly added CR or FD sample replaces one CC sample, so the pool gradually shifts toward more informative hard cases without changing its total size.

Each CC, CR, or FD sample is stored in a unified exemplar format. Each sample is stored as a retrievable exemplar containing the original counseling dialogue x, the gold label y, and type-specific auxiliary fields when needed. For CC samples, we store only the text and label (x, y); for CR samples, we store both the correct and incorrect rationales to provide contrastive guidance; for FD samples, we store an error diagnosis to surface likely failure causes and boundary-confusing patterns. This unified sample representation enables the retriever to return heterogeneous yet complementary exemplars.

At test time for a new query x_q, we retrieve the top-k exemplars from the exemplar pool and include them as in-context demonstrations. In our experiments, we fix $k = 64$. We further apply type-weighted retrieval to prioritize informative CR and FD samples during inference.

3.4. Curriculum-based Rule Construction and Refinement

We refine an executable rule set in a label-aware manner to mitigate rule bias under severe class imbalance. Besides the learning-value types derived from contrastive typing, we additionally condition rule induction on the gold risk label. This design is motivated by the asymmetric utility of errors in risk triage: under low high-risk prevalence, unconstrained aggregation across labels tends to overrepresent majority (low-risk) regularities, yielding conservative constraints that can suppress minority detection. By separating high-risk (y=1) and low-risk (y=0) evidence during refinement, we encourage the rule refiner to extract label-specific decision cues and to preserve high-risk sensitivity while retaining low-risk exclusion criteria.

Formally, we split the constructed samples by gold label, yielding six partitions:

$$\mathcal{S} = \{\text{CC-0},\ \text{CC-1},\ \text{CR-0},\ \text{CR-1},\ \text{FD-0},\ \text{FD-1}\}.$$

Within each label-specific partition, we further organize instances by a correctness axis that indicates whether the corresponding inference configuration predicts the gold label correctly. To avoid ambiguity with class polarity, we denote this axis as correct and incorrect. Correct instances provide higher-reliability evidence for initial constraint formulation, whereas Incorrect instances expose boundary confusions and missing conditions that are informative for subsequent rule revision.

We adopt a two-level curriculum schedule that is both label-aware and difficulty-aware, and proceeds in an explicit easy-to-hard, as shown in Figure 3. Here, difficulty is determined by the correctness outcomes of the two inference configurations. Samples for which both configurations are correct are treated as easy (CC), those for which only one configuration is correct are treated as intermediate (CR), and those for which both configurations are incorrect are treated as hard (FD). Within each gold-label group, evidence is further organized by correctness, and later stages incorporate less separable cases to revise and qualify previously induced constraints.

For each label group (y=0 and y=1), refinement starts from more discriminative and reliable evidence and progressively incorporates less separable evidence to amend previously induced constraints. Specifically, we introduce CC subsets first, then CR subsets, and finally FD subsets. CC cases are easy because both configurations are correct, CR cases are intermediate because only one is correct, and FD cases are hard because both are incorrect. This progression stabilizes early rule formation, while later stages emphasize Incorrect cases to introduce disambiguation constraints and exception constraints that improve robustness on borderline dialogues.

To control partition imbalance, we sample up to a predefined upper bound from each partition and process instances in mini-batches of 25.

Let Rule_t denote the rule set at step t. At each curriculum step, the rule-update prompt takes as input (i) the current rule set and (ii) a mini-batch of typed samples from the current subset. The LLM

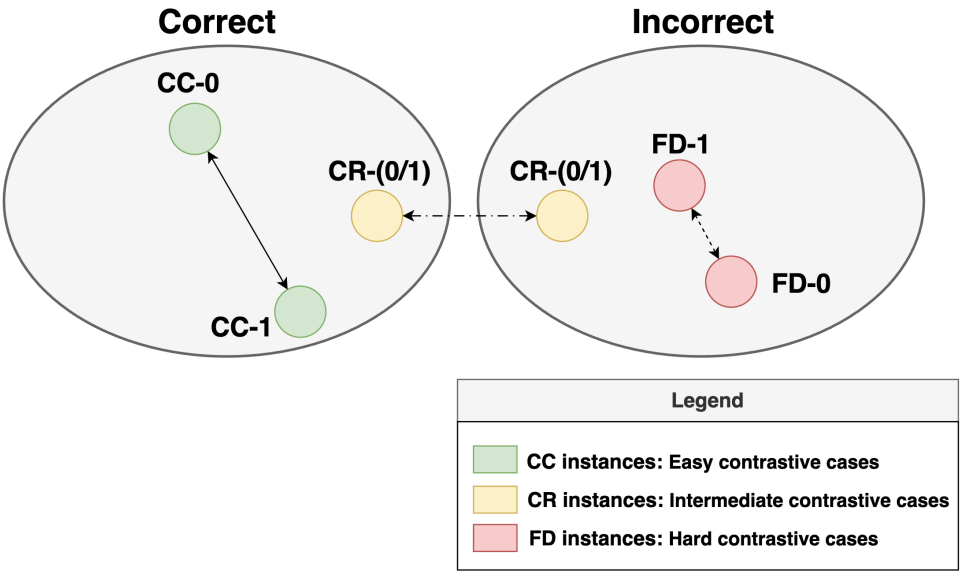


Figure 3: Label-aware easy-to-hard curriculum for rule refinement.

then regenerates the complete rule set in natural language and organizes it into four parts: high-risk indicators, low-risk indicators, false-positive prevention, and false-negative prevention. Given newly proposed constraints $\mathrm{New_Rule}_t$ summarized from the current batch, we update:

$$\mathrm{Rule}_t = \mathrm{Refine}\big(\mathrm{Rule}_{t-1} \cup \mathrm{New_Rule}_t\big). \quad (5)$$

We implement rule updates by prompting the LLM with the previous rule set and the newly introduced typed samples, then asking it to regenerate a revised full rule set. The updated rule set removes redundant constraints, merges overlapping rules, and relaxes overly specific formulations to improve generalizability.

3.5. Type-Weighted Exemplar Retrieval and Rule Injection

At test time, DDJR performs RA-ICL inference with type-weighted exemplar retrieval and full injection of the refined rule set. Given a query text x_q, each case s in the exemplar pool has type $\tau(s) \in \{\mathrm{CC}, \mathrm{CR}, \mathrm{FD}\}$ and similarity score $\mathrm{sim}(x_q, s)$. We apply a type-dependent weight, $w_{\tau(s)}$, to the similarity score:

$$\mathrm{score}(x_q, s) = \mathrm{sim}(x_q, s) \cdot w_{\tau(s)} \quad (6)$$

We select the top-k samples according to $\mathrm{score}(x_q, s)$ as in-context demonstrations, where higher weights emphasize CR and FD to surface discrepancy and failure cues for borderline inputs.

We assign larger type weights to CR and FD cases so that discrepancy and failure-oriented exemplars are prioritized during retrieval and injects the full refined rule set Rule_t into the prompt as explicit decision guidance to encourage consistent and evidence-aligned predictions.

4. Experiments

4.1. Implementation

Experiments were conducted on an annotated corpus of online student counseling records, consisting of 6,481 records labeled by trained annotators. All records were anonymized prior to use, with personally identifiable information removed. We adopted the binary risk setting of low-risk versus high-risk, where high-risk corresponded to high-acuity suicide risk and low-risk corresponded to low-acuity suicide risk.

The dataset was divided into training, validation, and test sets with the ratio of $8:1:1$, stratified to preserve label proportions across splits. The resulting labels were imbalanced, with high-risk samples accounting for approximately 7% of the corpus.

We used GPT-4.1 (Achiam et al., 2023) as the backbone model for all LLM-based methods and compared several prompting-based inference settings under a shared retrieval setup. Unless otherwise specified, RA-ICL retrieved $k = 64$ semantically similar labeled exemplars based on vector similarity for in-context prompting. In all experiments, the type weights were set to $w_{CC} : w_{CR} : w_{FD} = 1 : 1.5 : 1.5$. All methods were assessed with the same backbone model and retrieval setting, and we reported accuracy (ACC), high-risk recall, balanced accuracy (BACC), and Matthews Correlation Coefficient (MCC).

Method	ACC	MCC	Recall	BACC
Zero-shot learning	0.643	0.316	**1.000**	0.808
Few-shot learning (Random)	0.764	0.402	0.978	0.863
ExRAG (Kermani et al., 2025)	0.843	0.453	0.889	0.864
CoE (Lee et al., 2023)	0.737	0.376	0.978	0.848
Fine-tuning (Achiam et al., 2023)	**0.944**	0.561	0.578	0.775
DDJR (Ours)	0.915	**0.583**	0.867	**0.893**

Table 1: Comparison with different methods under the same dataset split, backbone model, and evaluation metrics.

Sample usage	ACC	MCC	Recall	BACC
CC (Kermani et al., 2025)	0.843	0.453	**0.889**	0.864
CC+FD	**0.923**	0.415	0.467	0.712
CC+CR	0.889	0.523	0.867	**0.879**
CC+FD+CR	0.905	**0.546**	0.844	0.877

Table 2: Impact of FD and CR on sample effectiveness.

All prediction prompts used a shared task-definition template and required the model to output a structured JSON object containing a binary label and a free-text rationale. The zero-shot setting used only the task instruction and output schema, whereas the RA-ICL setting prepended retrieved demonstrations in the same output format. At test time, DDJR made the final prediction with RA-ICL using the refined exemplar pool as in-context demonstrations and the refined rule set as explicit decision guidance.

4.2. Comparison with Baseline Methods

We adopt a hierarchy of increasingly competitive baselines to assess gains from generic prompting, contextual alignment, and iterative self-improvement. We also include a fine-tuned GPT-4o (Achiam et al., 2023) model for reference; without cost-sensitive tuning, fine-tuning may optimize aggregate accuracy while under-detecting the minority high-risk class, leading to lower Recall and BACC despite high ACC. The zero-shot setting relies only on task instructions and the target record, and is therefore susceptible to confusing transient emotional intensity with sustained psychological risk. Few-shot prompting improves label discrimination but can remain unstable due to sensitivity to the selected in-context examples. Building on RA-ICL, DDJR further refines the exemplar pool and rule set before test-time inference.

4.3. Structured Sample Usage

We treat ExRAG (Kermani et al., 2025) as the CC-only variant, where CC refers to standard consistently-correct exemplars. Its exemplar pool contains only text-label exemplars, without the CR (contrastive rationalization) or FD (failure-diagnostic) auxiliary fields used in DDJR.

Method	ACC	MCC	Recall	BACC
CC w/ CoE (Lee et al., 2023)	0.838	0.447	0.889	0.862
CC Non-curriculum strategy	**0.918**	**0.501**	0.667	0.802
CC CBRR	0.827	0.453	**0.933**	**0.876**

Table 3: Comparison of curriculum-based rule refinement (CBRR) starting from the same CC-only instance set.

We examine how different combinations of disagreement-derived sample types affect performance. In this comparison, only the included sample types are varied, while the backbone model, retrieval setting, and downstream inference procedure are kept unchanged, and no additional rule-update rounds are performed.

We compare four settings: (i) CC-only; (ii) CC+FD; (iii) CC+CR; and (iv) CC+FD+CR. This comparison examines whether different disagreement-derived sample types provide complementary benefits. Table 2 indicates that the contribution of typed samples is type-dependent.

Adding FD without CR reduces high-risk recall substantially, from 0.889 to 0.467, and also lowers BACC from 0.864 to 0.712, although ACC increases from 0.843 to 0.923. This suggests that many FD cases primarily emphasize correcting over-sensitive false-positive decisions; in the absence of contrastive guidance, such signals may inadvertently suppress high-risk predictions.

Incorporating CR improves the overall trade-off in MCC and BACC while largely preserving high-risk recall relative to CC-only. This supports the role of CR in providing contrastive evidence where the same input yields different outcomes across prompting configurations thereby encouraging alignment with more appropriate evidence usage.

Combining CR and FD yields the most balanced overall performance, suggesting that CC, CR, and FD provide complementary supervision under class imbalance: CC supplies stable but conservative signals, CR anchors decision boundaries via contrastive instances, and FD complements these signals by summarizing persistent failure patterns that remain even with retrieval augmentation.

4.4. Curriculum-Based Rule Refinement

We evaluate whether curriculum-based rule refinement (CBRR) is necessary for robust risk detection under severe class imbalance. For clarity, both "w/o CBRR" and "w/ CBRR" start from the same CC-only instance set, and differ only in whether the curriculum schedule is applied during rule refinement. We compare two settings: (i) without CBRR, where rule refinement is performed on the

same exemplar pool using a non-curriculum strategy (i.e., removing the easy-to-hard, label-aware progression); and (ii) with CBRR, which follows the proposed curriculum schedule. In both settings, the backbone model, retrieval pipeline, and evaluation protocol are kept identical; the only difference is whether the curriculum schedule is applied during rule refinement.

As shown in Table 3, removing CBRR significantly weakens the model's ability to identify high-risk samples. In contrast, incorporating CBRR better balances detecting high-risk samples against avoiding false alarms on low-risk ones, leading to the highest BACC of 0.876. CBRR improves minority sensitivity, yielding higher Recall and BACC, but also increases false positives, resulting in lower ACC. This trade-off is preferable in practical screening, where missing high-risk samples is much more costly than making errors on low-risk samples.

4.5. Ablation Study and Qualitative Analysis

Method	ACC	MCC	Recall	BACC
DDJR (Ours)	**0.915**	**0.583**	0.867	**0.893**
w/o Exemplar pool refinement	0.777	0.404	**0.956**	0.859
w/o Rule refinement	0.905	0.546	0.844	0.877

Table 4: The ablation study of DDJR.

To evaluate the contributions of the two key design choices in DDJR, we conduct an ablation study by removing: (i) exemplar pool refinement, which constructs the types of disagreement-derived instances via dual-inference comparison, and (ii) rule refinement, which refines the rule set following the proposed curriculum schedule. All variants are evaluated under the same LLM, prompt skeleton, and retrieval setting; the only changes are the removal of the corresponding component.

Table 4 reports the results of the ablation study. Removing contrastive typing (i.e. without exemplar pool refinement) leads to a substantial degradation in ACC and MCC, indicating that the construction of the three sample types is critical for producing informative supervision signals and for stabilizing decision boundaries. In contrast, removing rule refinement (without rule refinement) yields a smaller but consistent drop across MCC, Recall, and BACC, suggesting that the curriculum schedule contributes to maintaining balanced performance under class imbalance by refining rules in a structured manner. Collectively, the full DDJR configuration achieves the best performance, demonstrating that both contrastive typing and rule refinement are necessary to obtain robust counseling risk classification. Although DDJR does not achieve the highest high-risk recall, it achieves the best ACC (0.915), MCC (0.583), and BACC (0.893), suggesting that it offers a more balanced trade-off between sensitivity to high-risk cases and overall classification robustness.

Example Analysis

Text:
…But lately, I've started to strongly feel like I want to die. It's not like anything in particular happened—***my relationships and grades at school are decent—but I still feel like I want to die. I can't talk about this with my parents,*** teachers, or friends. Because I'm known as the "cheerful one," if I told people something like this, they'd worry about me. I can't bring myself to say "I want to die," not even as a joke. Why have these suicidal thoughts suddenly come up...

Ground-truth label: Low-risk

Method: Zero-shot learning
Result: High-risk
Reason: There have been ***suicide suggestive remarks from the parent***, such as "Shall we die together?" and "There's no point in living," ***and the person seeking help is under severe stress related to school, future plans, relationships, and the home environment;*** therefore, the suicide risk is assessed as high.

Method: DDJR
Result: Low-risk
Reason: The suicidal ideation ("strongly feel like I want to die") is clearly expressed, but there is no mention of a specific suicide plan, intent to act, self-harm behavior, or an imminent sense of crisis. The main complaint remains confusion about why these feelings arose and the conflict of being unable to seek help. ***Based on the [LowRisk Indicators] [False Positive Fix] rule, the suicide risk is judged to be low at this time.***

Figure 4: Case study.

Figure 4 presents a representative qualitative example illustrating how the proposed components mitigate typical false positives. The record describes a student's distress related to schooling and family pressure, and includes a parent's statement about "dying together". Our method predicts low-risk by attributing the suicidal wording to the parent rather than the student, and by identifying contextual cues indicating agency and forward-looking intent (e.g., a desire for independence and plans to continue living). In contrast, the baseline RA-ICL setting misclassifies the record as high-risk, overemphasizing death-related keywords and failing to account for speaker attribution and broader discourse context. This case suggests that contrastive typing supplies contrastive and failure-oriented exemplars that surface such misinterpretations, while the proposed rule set helps regularize the decision boundary by discouraging keyword-triggered overgeneralization. Together, these com-

ponents promote a more faithful interpretation of emotionally charged language and reduce false positives in risk triage.

5. Conclusion and Limitations

We propose DDJR, an iterative parameter-free refinement framework for imbalanced counseling risk classification that does not update model parameters. The method improves retrieval-augmented inference by refining the exemplar pool and decision rules through disagreement signals derived from zero-shot and RA-ICL predictions. Experimental results showed that DDJR achieved the best overall balance among ACC, MCC, Recall, and BACC, indicating more robust and reliable performance under severe class imbalance.

Several limitations should also be noted. First, the recall of DDJR indicates that some high-risk samples may still be missed. Second, the proposed refinement is conducted offline on labeled training data, while test-time use only relies on the finalized exemplar pool and refined rule set for inference. Third, the study is based on counseling records from a single institution, and the generalizability of the learned retrieval patterns and induced rules to other populations and settings remains unclear. Future work will examine external validation, cross-institutional transfer, and the stability of the refined rule set under distribution shift.

6. Acknowledgements

This work has been approved by the Institutional Review Board of our graduate school: UT-IST-RE-231102. This work was partially financially supported by JST ASPIRE Program, Japan, Grant Number JPMJAP2303 and The Toyota Foundation, Co-Creating New Society with Advanced TechnologiesGrant Number: D24-ST-0006.

Alaa A. Abd-Alrazaq, Asma Rababeh, Mohannad Alajlani, Bridget M. Bewick, and Mowafa Househ. 2020. Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 22(7):e16021.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, and J. Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive Science*, pages 179–211.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Mental Health*, 4(2):e19.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *ICML*, pages 3929–3938.

Y. Jiang, Y. Xiong, Y. Yuan, C. Xin, W. Xu, Y. Yue, and L. Yan. 2025. Pag: Multi-turn reinforced llm self-correction with policy as generative verifier. *arXiv preprint arXiv:2506.10406*.

Arshia Kermani, Veronica Perez-Rosas, and Vangelis Metsis. 2025. A systematic evaluation of LLM strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. RAG. pages 172–180.

Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. 2023. Chain of empathy: enhancing empathetic response of large language models based on psychotherapy models. *arXiv preprint arXiv:2311.04915*.

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. *NeurIPS*, pages 9459–9474.

Wenjie Li, Tianyu Sun, Kun Qian, and Wenhong Wang. 2024. Optimizing psychological counseling with instruction-tuned large language models. *arXiv preprint arXiv:2406.13617*.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y. Zhao. 2024. Dr.icl: Demonstration-retrieved in-context learning. *Data Intelligence*, pages 909–922.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *NeurIPS*, pages 46534–46594.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *CLPsych*, pages 39–44.

Calvin Mercer. 2022. Mental and spiritual health needs of cognitively enhanced people: a therapeutic and spiritual care model for responding. *Religions*, page 701.

Sungjoon Park, Donghyun Kim, and Alice Oh. 2019. Conversation model fine-tuning for classifying client utterances in counseling dialogues. In *NAACL-HLT*, pages 1448–1459.

Michiel Rauws, John Quick, and Nancy Spangler. 2019. X2 AI Tess: working with AI technology partners. *The Journal of Employee Assistance*, 49(1):22–25.

Hao Shen, Zihan Li, Minqiang Yang, Minghui Ni, Yongfeng Tao, Zhengyang Yu, Weihao Zheng, Chen Xu, and Bin Hu. 2024. Are large language models possible to conduct cognitive behavioral therapy? In *BIBM*, pages 3695–3700.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. *NeurIPS*, pages 8634–8652.

Anand N. Vaidyam, Hannah Wisniewski, John D. Halamka, Matcheri S. Kashavan, and John B. Torous. 2019. Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.

World Health Organization. 2023. Depressive disorder (depression).

Z. Wu, Q. Zeng, Z. Zhang, Z. Tan, C. Shen, and M. Jiang. 2024. Large language models can self-correct with key condition verification. In *EMNLP*, pages 12846–12867.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. STaR: Bootstrapping reasoning with reasoning. In *NeurIPS*, pages 15476–15488.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *npj Digital Medicine*, page 46.

A. Example Rule Set

The rules were originally generated in Japanese and are shown here in English for readability.

High-risk indicators. Specific suicide methods, suicide attempts, overdose (OD), wrist-cutting, use of means such as knives, suicide preparation/planning, a suicide note or will, or a history of emergency transport. If suicidal intent or action is clear, classify the case as high risk.

Low-risk indicators. Expressions such as “I want to die, but I do not want to die,” “Small things make me want to die,” or “I keep going back and forth between wanting to live and wanting to die,” when they reflect only ambivalence or temporary suicidal ideation without any concrete action or plan. In such cases, classify the case as low risk.

False-positive prevention. Even if phrases such as “I want to die” or “I want to disappear” appear, do not classify the case as high risk if there is no specific plan, method, intent, or self-harm behavior, and the text mainly expresses distress or help-seeking.

False-negative prevention. Even if phrases such as “I want to die” or “I want to disappear” are indirect, classify the case as high risk if they are accompanied by suicide preparation, planning, self-harm, suicide attempts, a suicide note, or strong hopelessness and urgency.

B. Cost and Scalability Discussion

DDJR introduces additional overhead compared with zero-shot and standard RA-ICL baselines, primarily because it uses longer prompts and an offline multi-round refinement procedure before final test-time prediction. Since DDJR does not update model parameters, its cost is better understood as online prediction cost plus offline refinement overhead rather than training cost.

Under our setup, all three methods require only one LLM call per test instance, but their prompt sizes differ substantially: approximately 1,000 characters for zero-shot, around 25,000 for standard RA-ICL, and slightly above 50,000 for DDJR. This corresponded to an estimated per-instance input cost of about \$0.003, \$0.04–\$0.05, and \$0.06–\$0.09, respectively. In addition, DDJR incurred a one-time offline refinement cost of approximately \$318 prior to final test-time inference. These observations suggest that the main overhead of DDJR comes from the one-time offline refinement stage, while deployment only uses the finalized exemplar pool and refined rule set for inference.

Context-Aware SNOMED CT Entity Linking for Clinical Text

Provia Kadusabe[1], **Demian Gholipour Ghalandari**[2], **Lauren Cassidy**[2], **Jack Boylan**[2], **Chris Hokamp**[2], **Abhishek Kaushik**[1], **Fiona Lawless**[1]
[1]Dundalk Institute of Technology, Ireland
[2]Quantexa Ltd., Dublin, Ireland
{provia.kadusabe, abhishek.kaushik, fiona.lawless}@dkit.ie
{demiangholipour, laurencassidy, jackboylan, chrishokamp}@quantexa.com

Abstract

Mapping free-text mentions in clinical notes to standardized terminologies such as SNOMED CT is essential for large-scale secondary use of electronic health records, but remains challenging due to linguistic variability, under-specified annotation guidelines, term ambiguity, and ontology scale. This work presents a two-stage entity linking pipeline that combines span detection with context-aware concept linking and evaluates it on the SNOMED CT Entity Linking Challenge dataset. Our work builds upon the SNOMED CT entity linking challenge (Davidson et al., 2025), resulting in a fully open-source system. To our knowledge, this is the first end-to-end open-source system for this task. For span detection, we compare multiple neural architectures together with dictionary-based matching. For concept linking, we adopt a context-aware bi-encoder, and construct a multi-source knowledge base enriched with context derived from the SNOMED CT ontology. Finally, we implement an agentic re-ranker and test the effectiveness of LLM-backed re-ranking with access to annotation guidelines. In contrast to findings from the original shared task submissions, we show that context is important for optimal performance, and that agentic re-ranking with a state-of-the-art LLM only marginally improves overall performance, suggesting that the current benchmark may be approaching its practical ceiling. This work provides the first fully open-source, reproducible system for SNOMED CT entity linking, offering a foundation for future research and practical deployment.

Keywords: Clinical Entity Linking, SNOMED CT, Clinical NLP, Medical Concept Disambiguation, Concept Normalization

1. Introduction

Clinical notes in Electronic Health Records (EHRs) contain rich information about patients and treatment protocols, but much of this data is unstructured, making it difficult to analyze at scale. Mapping free-text mentions in clinical notes to standardized medical terminologies such as SNOMED CT [1] is a crucial step toward enabling large-scale secondary use of clinical data, facilitating clinical decision support, pharmacovigilance, and biomedical research. This task is commonly referred to as clinical entity linking or concept normalization (French and McInnes, 2023), which requires both accurately identifying spans of text that refer to medical entities and correctly mapping them to the appropriate concept identifiers. Figure 1 illustrates the task: free-text mentions in clinical notes are mapped to their corresponding SNOMED CT concepts.

Clinical entity linking presents several challenges. Clinical narratives make pervasive use of abbreviations and shorthand that may have multiple meanings (such as RA: Rheumatoid Arthritis vs. Right Atrium), requiring disambiguation during linking (Davidson et al., 2025). Concept granularity and semantic overlap further complicate candidate selection, as closely related concepts may differ only in subtle ways (such as “chest pain” vs. “anterior chest wall pain”). Span boundary detection poses additional difficulty, as modifiers, nested expressions, and overlapping mentions can lead to multiple plausible spans, each mapping to a different clinical concept. Finally, the target concept space is large; clinical terminologies such as SNOMED CT contain hundreds of thousands of concepts, yet annotated datasets typically cover only a small fraction, producing a long-tail distribution that requires systems to generalize to rare and previously unseen concepts at inference time (Davidson et al., 2025).

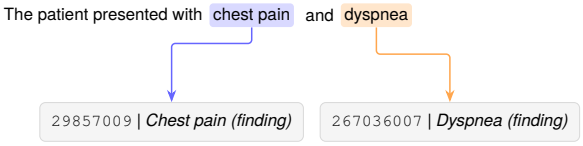


Figure 1: SNOMED CT entity linking: mentions in clinical text (highlighted) are linked to standardized concepts in the SNOMED CT ontology

The recent SNOMED CT Entity Linking Challenge (Davidson et al., 2025) provided a large, publicly available benchmark for this task, derived from de-identified discharge summaries in MIMIC-IV-Note (Johnson et al., 2023). The dataset includes about 75,000 manually annotated mentions spanning three high-level SNOMED CT subhierarchies: `Findings`, `Procedures`, and `Body Structures`. This benchmark has driven recent

[1] https://www.snomed.org

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 290–299
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

research on entity linking in clinical text. However, most existing top systems from the challenge largely fall into two paradigms: (i) highly engineered dictionary-based approaches that rely upon lexical matching and heuristics, or (ii) neural retrieval-based pipelines that primarily treat linking as string-to-concept matching, with limited use of broader document context.

In this work, we propose a two-stage entity linking pipeline that integrates heterogeneous span detection methods with context-aware retrieval. In the first stage, we compare three neural architectures for span detection with different inductive biases and model sizes: `GatorTron`, a large clinical language model fine-tuned with QLoRA (Yang et al., 2022); `BioClinicalBERT` (Alsentzer et al., 2019), augmented with a conditional random field decoding layer; and `GLiNER-BioMed` (Yazdani et al., 2025), a zero-shot, span-based, type-conditioned model. In the second stage, we adopt a contextual bi-encoder retrieval approach using `KRISSBERT` (Zhang et al., 2022), which encodes both mentions and knowledge base entries with surrounding document context rather than treating mentions as isolated strings. We construct a structured multi-source knowledge base that combines official SNOMED CT terminology, contextualized mention prototypes from the training data, OMOP/Athena[2] synonyms aligned to SNOMED concepts, and ontology-derived pseudo-sentences constructed from SNOMED relationship triples such as *is-a*, *finding site*, and *morphology* relations. This enables the model to leverage both local discourse cues and structured medical knowledge.

The key contributions of this work are:

- We propose and implement a fully reproducible, end-to-end framework for clinical entity linking.[3]
- We demonstrate that context-aware retrieval improves benchmark results over mention-only encoding, a finding not shown in previous SNOMED CT task submissions.
- We provide a thorough evaluation and detailed error analysis that reveal key challenges in clinical entity linking, especially with respect to the current SNOMED CT dataset.

2. Related Work

2.1. Background

Clinical entity linking aims to map text spans in clinical narratives to standardized medical concepts. As of early 2026, SNOMED CT serves as the primary knowledge resource for this task. The knowledge base contains over 350,000 clinical concepts organized in a poly-hierarchical structure, with each concept defined by a fully specified name, synonyms, and semantic relationships such as `finding site` and `causative agent`. In addition to SNOMED CT, clinical entity linking systems often rely on other complementary knowledge resources. The UMLS Metathesaurus (Unified Medical Language System[4]) (Bodenreider, 2004) integrates nearly 200 source vocabularies including SNOMED CT, ICD-10, MeSH, and RxNorm, into a shared concept space containing over 3 million concepts and more than 17 million unique concept names. UMLS provides cross-vocabulary mappings and synonym sets that are widely used for training entity linking models (Liu et al., 2021; Sung et al., 2020). The OMOP/Athena repository similarly provides standardized mappings across clinical coding systems and additional terminology.

2.2. Approaches for Clinical Entity Linking

Early approaches to clinical entity linking relied heavily on rule-based and dictionary-based systems such as MetaMap (Aronson, 2001) and cTAKES (Savova et al., 2010), which match text spans to terminology entries using curated lexicons and handcrafted heuristics. While these methods can achieve high precision for well-defined expressions, they struggle with linguistic variability, abbreviation ambiguity, and unseen terminology, and often require extensive manual engineering.

With the advent of pre-trained base models (Devlin et al., 2019), neural approaches have become dominant. Many recent systems frame entity linking as a retrieval problem, where mentions and concepts are embedded into a shared vector space and matched using cosine similarity. Encoder models such as SapBERT (Liu et al., 2021), BioSyn (Sung et al., 2020), and PubMedBERT (Gu et al., 2021) have been widely adopted for biomedical entity linking. These bi-encoder models are efficient at scale but typically encode mentions with little or no broader document context.

Prior work adopt a two-stage architecture (Kartchner et al., 2023; Gallego et al., 2024; French and McInnes, 2023; Sung et al., 2020): a bi-encoder first retrieves a small set of high-probability candidate concepts, then a context-aware re-ranker re-scores these candidates using a cross-encoder or language model that conditions on a textual window around the mention.

[2] https://athena.ohdsi.org/

[3] https://anonymous.4open.science/r/snomed-ct-entity-linking-project-E62B

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

Recent context-aware biomedical entity linking models such as KRISSBERT (Zhang et al., 2022) are designed to capture richer contextual information than standard bi-encoders during retrieval. However, to the best of our knowledge, these context-first retrieval models have not been systematically evaluated on SNOMED CT entity linking in clinical text.

2.3. SNOMED CT Entity Linking Challenge

The SNOMED CT challenge provided a large-scale human-annotated dataset in which spans of text are labeled with their corresponding SNOMED CT concepts (Davidson et al., 2025). The top performing systems of the challenge employed different approaches. The first-place solution in the SNOMED CT Entity Linking Challenge employed a purely dictionary-based approach (Davidson et al., 2025). During inference, documents were split into sections and mentions were detected via regex matching.

The second-place solution (Kulyabin et al., 2024) employed a two-stage pipeline. First, an ensemble of six fine-tuned `BiomedBERT-large` models detected medical entities. Then `SapBERT` was used for concept linking. Additional post-processing included dictionary filtering, section removal, and confidence thresholds. However we note discrepancies between reported methods and released implementations; for example, the paper mentions a re-ranking step that is not present in the available implementation.

The third-place solution used an LLM-based approach for both entity recognition and classification. Two fine-tuned `Mistral` models detected entities on different chunk sizes, with dictionary-based post-processing. In the second stage, a Faiss index (Johnson et al., 2019) retrieved candidate concepts, and a third `Mistral` model selected the final concept. The third place uses a small window of context but only in the 3rd stage for re-ranking.

Our work takes inspiration from these solutions but also addresses some of the issues we identified, such as discrepancies between reported methods and released implementations, and unclear evaluation methodology which is not possible to fully reproduce in its presented form. In this work, we provide a fully reproducible system that additionally incorporates context at the retrieval stage, an approach not implemented by any of the challenge submissions and cleanly separate span detection evaluation from concept linking evaluation.

3. Methods

3.1. Data

We used version 1.1 of the SNOMED CT Entity Linking Challenge dataset.[5] The full dataset contains 272 annotated clinical notes with character-level entity annotations, each consisting of a `note identifier`, `start` and `end` offsets, and a SNOMED CT concept identifier. The original challenge was based on the May 2023 International Edition of SNOMED CT and restricted to three sub-hierarchies (`Clinical Findings`, `Procedures`, and `Body Structures`).[6] Table 1 reports summary statistics of the original challenge dataset. Due to availability constraints, we used the November 2025 International Edition of SNOMED CT in our experiments instead of the May 2023 version used in the challenge. This version mismatch resulted in 102 concepts being unavailable (due to deprecation or restructuring), affecting 611 annotations in total (409 in the training set and 202 in the test set).

Split	Notes	Annotations	Unique Concepts	Unseen in Train
Train	204	51,574	5,336	–
Test	68	23,234	4,082	1,288

Table 1: Summary statistics of the SNOMED CT Entity Linking Challenge dataset.

3.2. Entity Linking Pipeline

Our system follows a two-stage pipeline architecture for SNOMED CT entity linking in clinical text, as illustrated in Figure 2.

Stage 1: Span Detection We evaluate three neural architectures for entity span detection. `GatorTron` (Yang et al., 2022) a large clinical language model that we fine-tune using parameter-efficient LoRA adapters with 4-bit quantization. `BioClinicalBERT-CRF` which we build on `Bio_ClinicalBERT` (Alsentzer et al., 2019) by adding a conditional random field layer to better model dependencies between adjacent labels during decoding. Both `GatorTron` and `BioClinicalBERT-CRF` frame NER as a token classification problem using BIO tagging. `GLiNER-BioMed` (Yazdani et al., 2025) is a zero-shot span-based model that represents entity types as learnable embeddings and scores candidate text spans directly against these representations.

[5] The dataset is publicly available on PhysioNet and is derived from the MIMIC-IV-Note corpus of de-identified discharge summaries from Beth Israel Deaconess Medical Center (Hardman et al., 2025).

[6] Further details about the challenge are described in (Davidson et al., 2025).

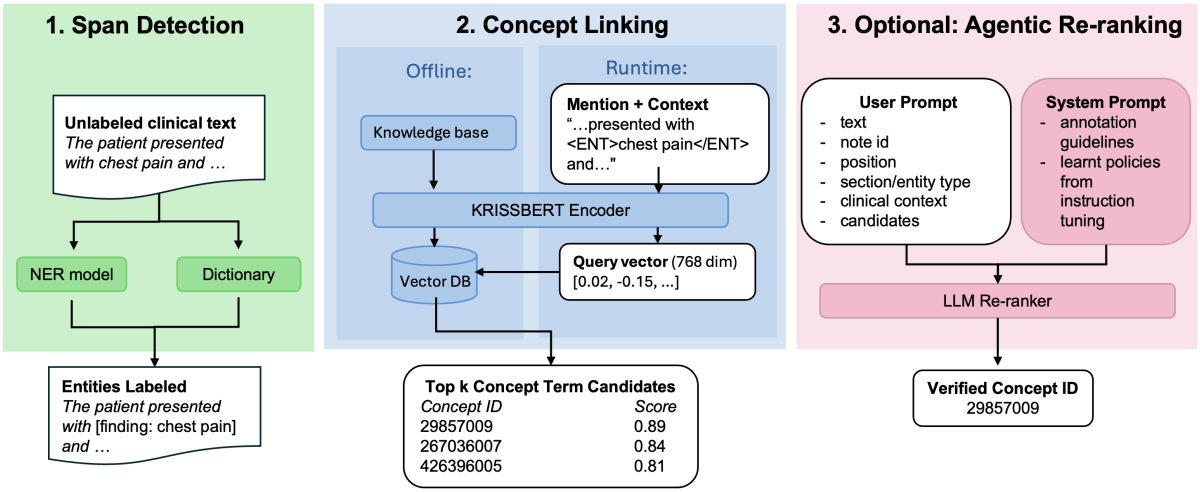


Figure 2: Our three-stage SNOMED CT entity linking pipeline: (1) span detection via neural NER (`GatorTron`, `BioClinicalBERT-CRF`, or `GLiNER`) and dictionary matching, (2) bi-encoder concept retrieval using `KRISSBERT` against a multi-source knowledge base (SNOMED terms, training prototypes, OMOP synonyms, and ontology pseudo-sentences), and (3) optional agentic re-ranking.

All models predict three SNOMED CT semantic categories: procedure, finding, and body structure. We additionally employ dictionary-based detection using an Aho-Corasick automaton (Aho and Corasick, 1975) constructed from surface forms in the training set. Terms with training precision below 0.50 are blacklisted to filter ambiguous matches. Predictions are post-processed by removing spans in medication sections or section headers, and filtering stopwords. Overlapping spans are resolved by retaining the longest match. The neural and filtered dictionary predictions are then merged and deduplicated to obtain candidate spans for linking.

Stage 2: Concept Linking Stage 2 takes the entity spans from stage 1 as input and maps them to their respective SNOMED CT concept identifiers. We adopt a bi-encoder-style retrieval approach using `KRISSBERT` (Zhang et al., 2022), a biomedical entity linking model pre-trained on PubMed and UMLS. Mentions and candidate terms are wrapped with entity markers (`<ENT>` and `</ENT>`) and encoded into L2-normalized representations. For comparison, we also implement `SapBERT` as a baseline; key architectural and training differences between `SapBERT` and KRISSBERT are summarized in Table 2. We implement the `KRISSBERT` aligned packing by tokenizing left context, mention, and right context separately as shown in Figure 3 and truncating context symmetrically to preserve the marked mention span. The resulting embeddings are stored in a Qdrant[7] vector database and retrieved using cosine similarity.

The knowledge base integrates four sources: (1) SNOMED CT terms extracted from RF2 releases and filtered to include only the three relevant subset SNOMED sub-hierarachies; (2) training mention prototypes augmented with a surrounding context window; (3) OMOP/Athena vocabulary synonyms mapped to SNOMED concepts to expand lexical coverage; and (4) ontology-derived pseudo-sentences constructed from SNOMED relationship triples by incorporating *is-a* relations, finding sites, and associated morphology.

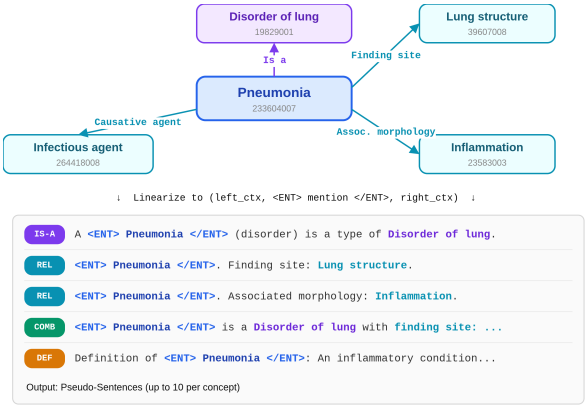


Figure 3: Pseudo-sentence construction from SNOMED CT relationship triples. Each concept's local ontology neighbourhood is linearized into up to 10 pseudo-sentences using four templates: IS-A (parent/class relationships), REL (single attribute relationships), COMB (a combined parent + key attributes in one sentence), and DEF (text definitions when available). Each pseudo-sentence is represented as (`left_ctx`, `<ENT>` mention `</ENT>`, `right_ctx`) and encoded with KRISSBERT for indexing in Qdrant.

At inference, each mention is encoded together

[7] https://github.com/qdrant/qdrant

Aspect	SapBERT	KRISSBERT
Backbone model	PubMedBERT	PubMedBERT
Training data	UMLS ontology (concept names and synonyms)	UMLS ontology + large unlabeled biomedical text (such as PubMed abstracts)
Input format	Entity name / synonym (no surrounding sentence)	Sentence with marked mention span (special entity markers)
Pooling	Mean pooling over token embeddings of the entity name	CLS representation of the contextualized mention
Training objective	Self-alignment via contrastive learning on synonyms	Knowledge-rich self-supervision + contrastive learning on contextualized mentions mined from text

Table 2: Differences between SapBERT and KRISSBERT

with its surrounding context window and entity markers. The query embedding is compared against the knowledge base using cosine similarity. Multiple matched surface forms mapping to the same concept are deduplicated by retaining only the highest-scoring match per concept ID, and the top-k ranked concepts after deduplication are returned.

Stage 3: Agentic Re-ranking (optional) We additionally implement an agentic re-ranker which has access to the top-k concepts from the concept linking step, and the annotation guidelines given to annotators [8]. The re-ranker implementation is motivated by the observation that correct candidates are often present in the top-k concepts retrieved from the knowledge base. The agent learns its re-ranking policy by sampling outputs from the training dataset. For each mention, the model receives: mention text and offsets, local note context (120 characters per side), and the top-5 linked candidates with enriched SNOMED metadata (FSN, semantic tag, embedding score, synonyms, definition, parent concepts, and defining relationships). The system prompt includes compressed annotation guidelines plus learned policies, capped at 20k characters for cost control.

The re-ranker produces one of three actions (`re-rank`, `drop`, `modify_span`). In the policy optimization step, the agent learns that the Stage 2 linker is already strong and learns conservative post-hoc guardrails to minimize harmful changes: (i) re-rank only mentions with top-1 score ≤ 0.986, (ii) keep top-1 changes only when the original top-1/top-2 score gap is ≤ 0.001, (iii) require `high` confidence for top-1 changes, and (iv) disable `drop` and non-top-1 reorders.

4. Experiments

We note that direct comparison with the top-performing systems from the SNOMED CT Entity Linking Challenge is not feasible for several reasons. First, the underlying SNOMED CT release file used in this work differs from the version used during the challenge, which affects the dataset. Second, while the challenge codebases are publicly available, the actual submission outputs are not, making it impossible to evaluate prior systems under identical conditions. Third, as discussed in Section 2, we found discrepancies between reported methods and released implementations for some solutions. For these reasons, we focus on evaluating our pipeline components in a controlled and reproducible setting rather than attempting potentially misleading comparisons with prior submissions.

For span detection, we experiment with the training dataset mapped to three entity types: *procedure*, *finding*, and *body structure*. We use a fixed 4-fold note-level split, holding out fold 0 for validation and using folds 1–3 for training. Long documents are split into overlapping chunks of 512 tokens with stride 128. We compare three architectures: (1) `UFNLP/gatortron-large` fine-tuned using QLoRA with 4-bit quantization and LoRA adapters (r=8, α=16, lr=2e-4); (2) `emilyalsentzer/Bio_ClinicalBERT` with a CRF layer for sequence labeling (lr=3e-5); and (3) `Ihor/gliner-biomed-large-v1.0`, a span-based model (lr=5e-6). All models are trained for up to 30 epochs with early stopping (patience=3)[9].

4.1. Evaluation Metrics

Span Detection Metrics We evaluate span detection using `strict`, `exact`, and `partial` matching criteria from `nervaluate`[10]. Strict matching requires exact span and entity-type agreement, exact matching ignores entity type, and partial matching awards partial credit for overlapping spans.

Concept Linking Metrics We evaluate concept linking performance using Recall@K and Mean Reciprocal Rank (MRR). Recall@K measures the pro-

[8] https://github.com/google/healthcare-text-annotation/blob/master/guidelines/defining-entity-categories.md

[9] All code is available at https://github.com/PROVIA1/snomed-ct-entity-linking-project

[10] https://github.com/MantisAI/nervaluate

portion of mentions for which the correct concept appears within the top-K predicted candidates.

Macro-Averaged Character-Level IoU The primary evaluation metric used in the in the SNOMED CT Entity Linking challenge is macro-averaged character-level Intersection-over-Union (IoU). For a given concept class c, let P^c_{char} be the set of all characters that fall within any predicted span assigned to c, and let G^c_{char} be the corresponding set of characters covered by gold spans of class c. The IoU for class c is defined as:

$$\text{IoU}_c = \frac{|P^c_{char} \cap G^c_{char}|}{|P^c_{char} \cup G^c_{char}|}.$$

Let $\mathcal{C}$ denote the set of all concept classes that appear in either the predicted annotations or the gold annotations. The overall score is obtained by averaging the per-class IoU values across this set:

$$\text{Macro-IoU} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{IoU}_c.$$

This metric is particularly challenging because it is highly sensitive to boundary errors and class frequency. As IoU is computed at the character level, even small span mismatches (such as missing modifiers or slight boundary shifts) can lead to substantial score reductions. Moreover, macro-averaging assigns equal weight to all concept classes, so rare concepts strongly affect the score.

5. Results

We evaluate our two-stage entity linking pipeline in two parts. First, we report end-to-end system performance. Second, we evaluate each stage performance.

5.1. End-to-End Performance

End-to-end evaluation of our system shown in Table 3 reveals that `KRISSBERT` consistently outperforms `SapBERT`. `KRISSBERT` + Dict uses the spans generated by the Stage-1 `GatorTron` + dictionary merge, whereas `KRISSBERT` and `SapBERT` use spans produced by GatorTron alone. Using spans produced by the Stage-1 dictionary merge yields further gains in end-to-end Recall and MRR.

Table 4 reports macro-averaged character IoU at different retrieval confidence score thresholds, where predictions are retained only if the model's top-1 score exceeds the given threshold. We observe that a retrieval confidence threshold of 0.98 yields the highest macro-averaged character IoU for both `KRISSBERT` and `KRISSBERT` + Dict.

Model	Recall@1	Recall@5	Recall@10	MRR
SapBERT	60.66	73.88	74.93	0.666
KrissBERT	69.56	74.93	75.43	0.720
KrissBERT + Dict	**72.65**	**78.46**	**78.99**	**0.752**

Table 3: End-to-End Concept Linking Performance

Model	None	≥0.95	≥0.98	≥0.99
SapBERT	0.3189	**0.3413**	0.3339	0.3271
KrissBERT	0.3429	0.3457	**0.3615**	0.2898
KrissBERT + Dict	0.3658	0.3686	**0.3869**	0.3154

Table 4: Macro-averaged Character IoU at Different Score Thresholds

5.2. Stage 1: Span Detection Performance

Table 5 summarizes overall model performance under strict, exact, and partial evaluation.

`GatorTron-Large` achieved the best overall F1 scores for all the metrics, while `BioClinicalBERT` obtained the highest recall under partial matching. `GLiNER-BioMed-Large` performed consistently lower overall.

Per-entity results (Table 6) show substantial variation across entity types. GLiNER-BioMed-Large performed best on *Finding* entities but showed lower performance on *Procedure* recognition. `GatorTron-Large` achieved the strongest results for *Procedure* entities, while *Body Structure* entities were the most challenging across models.

5.3. Stage 2: Concept Linking Perfomance

Table 7 reports performance of our system when evaluated on exact gold spans, `KRISSBERT` substantially improves Recall@1 over `SapBERT` (+11.6 points), confirming the benefit of contextual encoding for disambiguation.

5.4. Agentic Re-ranking Performance

Initially, we expected the agentic re-ranker, powered by a state-of-the-art LLM with access to the full annotation guidelines and rich SNOMED CT metadata, to recover a substantial portion of the near-miss errors identified in our analysis . However, the gains were surprisingly modest.

At 5,000 mentions (Table 8), the re-ranker intervenes on 12.9% of cases but retains only 49 effective top-1 changes, yielding +11 improved versus 4 degraded on gold spans. We did not scale to the full test set as the marginal gains did not justify the computational cost. The limited success of this approach may indicate that achievable performance on this dataset is approaching saturation,

Model	Strict			Exact			Partial		
	P	R	F1	P	R	F1	P	R	F1
GatorTron-Large	**0.751**	**0.750**	**0.750**	**0.770**	**0.768**	**0.769**	**0.824**	0.822	**0.823**
BioClinicalBERT	0.723	0.749	0.736	0.740	0.767	0.753	0.799	**0.828**	0.814
GLiNER-BioMed-Large	0.638	0.636	0.637	0.654	0.651	0.653	0.769	0.765	0.767

Table 5: Stage 1 NER Model Comparison across Evaluation Metrics

Model	Finding			Procedure			Body Structure		
	P	R	F1	P	R	F1	P	R	F1
GatorTron-Large	0.767	0.774	0.770	**0.774**	0.747	**0.760**	0.653	0.670	**0.661**
BioClinicalBERT	0.747	0.759	0.753	0.752	**0.765**	0.758	0.592	**0.680**	0.633
GLiNER-BioMed-Large	**0.775**	**0.777**	**0.776**	0.416	0.410	0.413	**0.663**	0.658	**0.661**

Table 6: Per-Entity-Type Performance (Strict Metric)

Model	Recall@1	Recall@5	Recall@10	MRR
SapBERT	78.99	96.20	97.56	0.867
KrissBERT	**90.58**	**97.57**	**98.21**	**0.937**
KrissBERT + Dict	90.23	97.46	98.11	0.935

Table 7: Concept Linking Performance: Linking-only evaluated on predictions with exact span matches to gold annotations.

with remaining errors driven less by insufficient reasoning and more by inherent ambiguity in both the span annotations and concept linking targets.

6. Error Analysis

We conducted a detailed error analysis on the test set predictions from our best-performing configuration (`GatorTron` NER + dictionary merge + `KRISSBERT` linking).

Of 24,555 predictions against 23,032 gold annotations, our pipeline achieved 80.5% exact span coverage of gold (18,543 matches). Among exact span matches, 90.2% (16,732) were linked to the correct SNOMED CT concept, while 9.8% (1,811) were linked to incorrect concepts. We categorize errors into three types: false negatives (missed mentions), false positives (spurious predictions), and concept linking errors.

False Negatives The pipeline missed 4,489 gold mentions (19.5% of annotations). Analysis revealed that 60.4% (2,713) of missed mentions had partial span overlap with predictions, indicating boundary detection issues rather than complete detection failures. Common boundary errors included spans differing by 1–3 characters (such as predicting “BLOOD Glucose” instead of “Glucose”). The remaining 39.6% (1,776) were completely undetected. These were predominantly generic clinical terms, such as “Evaluation procedure” (48 instances), “Recommendation to stop drug treatment” (25), and “Patient medication education ” (17).

False Positives The pipeline generated 6,012 spurious predictions not present in gold annotations. Common false positives included ambiguous short terms: “left/right” (81 instances), “BLOOD Glucose” (74), “Alert and” (59), and generic terms like “surgery,” “normal,” and “medications.”

Concept Linking Errors Among the 1,811 concept linking errors (correct span, wrong concept), we observed that 66.4% involved predictions within the same semantic category as the gold concept (such as both being procedures), indicating fine-grained disambiguation challenges rather than categorical errors. In particular, these errors occurred with high model confidence which suggests that embedding similarity alone is insufficient for distinguishing closely related concepts. Analysis of candidate rankings revealed that in 80.7% of linking errors, the gold concept appeared within the top-10 retrieved candidates. Table 9 shows the distribution of gold concept positions when the top-1 prediction was incorrect. Nearly half (49.0%) of errors had the gold concept at position 2, just below the selected prediction. An additional 25.0% had gold at positions 3–5. Only 19.3% of errors had the gold concept absent from the top-10 retrieved candidates.

7. Discussion

In this section, we discuss our findings, identify key performance bottlenecks, and outline implications for future research.

7.1. Context as the Key Differentiator in Concept Linking

On exact span matches, `KRISSBERT` achieves a Recall@1 of 90.58%, compared to 78.99% for `SapBERT`, an absolute improvement of 11.6 percentage points (Table 7). This gap is consistent across retrieval depths, with `KRISSBERT` maintaining its advantage at Recall@5 (97.57% vs. 96.20%) and

n	re-ranked	Effective top-1	Improved	Degraded	Unchanged	Non-gold	Δ Recall@1	Δ Macro-IoU	Tokens (in/out)	Cost (USD)
200	26	3	1	1	0	1	+0.0000	+0.000157	0 / 0 (cached)	0.00
500	88	9	4	1	0	4	+0.0002	+0.000456	700,462 / 13,338	2.30
5000	645	49	11	4	3	31	+0.0004	**+0.001062**	5,655,294 / 129,636	18.91

Table 8: Conservative re-ranker scaling results on the first n test mentions (offset 0). Non-gold counts indicate top-1 changes on predicted spans without a matching gold span at evaluation time.

Gold Position	Count	% of Errors
Position 2	888	49.0%
Position 3–5	452	25.0%
Position 6–10	121	6.7%
Not in candidates	350	19.3%

Table 9: Distribution of gold concept positions in candidate list when top-1 prediction is incorrect.

Recall@10 (98.21% vs. 97.56%). The performance difference can be attributed to a fundamental architectural distinction between the two models. `SapBERT` encodes mentions in isolation, while `KRISSBERT` incorporates local context, which is particularly important in clinical text where abbreviations and short expressions are common. This finding aligns with recent literature. Kartchner et al. (2023), in their systematic evaluation of nine biomedical entity linking models, identified the effective incorporation of context into linking decisions as a persistent gap across current methods. The SNOMED CT Challenge (Davidson et al., 2025) similarly emphasized the need to better leverage contextual information when training examples are scarce. The results confirm that the benefits of contextual encoding extend to the large, multi-hierarchical ontologies like SNOMED CT.

7.2. Span Detection as the Performance Bottleneck

Error propagation is a well-documented limitation of pipeline-based biomedical entity linking, where mistakes in mention detection impose an upper bound on linking performance (Noh and Kavuluru, 2021; Sarol et al., 2024; Kartchner et al., 2023). Our results confirm that span annotation is a major bottleneck in the SNOMED CT entity linking task. When correct spans are provided, KRISSBERT achieves 90.58% Recall@1 (Table 7), yet end-to-end mIoU reaches only 0.3869. This is further supported by our upper bound analysis (see Appendix A) where providing gold spans increases Recall@10 from 78.99% to 96.12%. This gap originates primarily from Stage 1, where our best NER model achieves a strict F1 of only 0.750, missing 19.5% of gold mentions. Of the missed mentions, 60.4% exhibit partial span overlap, indicating boundary errors rather than detection failures. The remaining 39.6% are dominated by implicit clinical concepts such as "Evaluation procedure" and "Patient medication education" and may reflect annotation ambiguity.

The three NER architectures we evaluated make different kinds of errors which reveals the challenging nature of this task. `GatorTron-Large` provides the most balanced performance overall (strict F1: 0.750), whereas `GLiNER-BioMed-Large` achieves the highest F1 (0.776) for Findings but performs poorly on Procedures (0.413). In contrast, `BioClinicalBERT` attains a strict F1 of 0.736 and the highest partial recall of all models (0.828) despite being approximately 80 times smaller than `GatorTron-Large`, indicating that smaller models can still be highly competitive for clinical NER. Together, these results show that no single architecture handles all entity types equally well, indicating potential for ensemble strategies that route predictions through the strongest model for each type.

The continued need to complement neural methods with dictionary-based approaches across different architectures suggests a genuine limitation of current span detection models or ambiguity in the task specification itself.

8. Conclusion

In this work, we presented a two-stage entity linking pipeline for mapping clinical text to SNOMED CT concepts, combining neural span detection with contextual bi-encoder retrieval. We also implemented an agentic re-ranker and showed that it does not substantially improve performance. Our results demonstrate that context-aware encoding via `KRISSBERT` substantially outperforms context-free alternative for concept linking, and that ontology-derived pseudo-sentences provide meaningful knowledge base enrichment, but specification, annotation quality, and evaluation metrics can likely be improved for future versions of this task. The proposed system is fully reproducible and adaptable to other clinical entity linking tasks.

9. Limitations

Several limitations should be acknowledged. First, our experiments use the November 2025 International Edition of SNOMED CT rather than the May 2023 edition used in the original challenge. This affects 102 deprecated or restructured concepts (611 annotations), potentially limiting comparability with

prior results and imposing a practical ceiling on our reported performance. This version mismatch highlights an underappreciated challenge for clinical NLP systems: ontology versioning. SNOMED CT releases new editions bi-annually, and any production system must contend with concept deprecation, merging, and restructuring over time.

Second, while our agentic re-ranker produces consistent positive gains under conservative guardrails, the improvements remain small relative to the computational cost, and the approach was not scaled to the full test set. Whether more sophisticated re-ranking strategies can yield larger gains remains an open question.

Third, our evaluation is limited to English-language clinical text from a single institution (Beth Israel Deaconess Medical Center). The generalizability to other clinical settings, documentation styles, and languages remains to be evaluated.

Finally, the macro-averaged character-level IoU metric, while providing a unified measure of system performance, conflates span detection and concept linking errors in ways that can obscure the contributions of individual pipeline components. Our separate evaluation of NER (via nervaluate) and linking (via Recall@k on exact spans) addresses this to some extent, but standardized component-level evaluation protocols would benefit the broader community.

10. Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions, which helped improve the quality of this manuscript.

The authors also acknowledge the support and collaboration of Quantexa, and in particular the Ireland-based Quantexa NLP group.

This research was partly funded through the CREATE-DkIT project, supported by the HEA TU-Rise program and co-financed by the Government of Ireland and the European Union through the Southern, Eastern & Midland Regional Program of the ERDF 2021-27 and the Northern & Western Regional Programme 2021–27.

11. References

Alfred V Aho and Margaret J Corasick. 1975. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340.

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.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Rory Davidson, Will Hardman, Guy Amit, Yonatan Bilu, Vincenzo Della Mea, Aleksandr Galaida, Irena Girshovitz, Mikhail Kulyabin, Mihai Horia Popescu, Kevin Roitero, et al. 2025. Snomed ct entity linking challenge. *Journal of the American Medical Informatics Association*, 32(9):1397–1406.

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.

Evan French and Bridget T McInnes. 2023. An overview of biomedical entity linking throughout the years. *Journal of biomedical informatics*, 137:104252.

Fernando Gallego, Guillermo López-García, Luis Gasco-Sánchez, Martin Krallinger, and Francisco J Veredas. 2024. Clinlinker: Medical entity linking of clinical concept mentions in spanish. In *International Conference on Computational Science*, pages 266–280. Springer.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Will Hardman, Mark Banks, Rory Davidson, Donna Truran, Nindya Widita Ayuningtyas, Hoa Ngo, Alistair Johnson, and Tom Pollard. 2025. SNOMED CT Entity Linking Challenge. *PhysioNet*. Version 1.1.0.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*. Version 2.2.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547.

David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie Mitchell. 2023. A comprehensive evaluation of biomedical entity linking models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14462–14478.

Mikhail Kulyabin, Gleb Sokolov, Aleksandr Galaida, Andreas Maier, and Tomas Arias-Vergara. 2024. Snobert: A benchmark for clinical notes entity linking in the snomed ct clinical terminology. In *International Conference on Pattern Recognition*, pages 154–163. Springer.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4228–4238.

Jiho Noh and Ramakanth Kavuluru. 2021. Joint learning for biomedical ner and entity normalization: encoding schemes, counterfactual examples, and zero-shot evaluation. In *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–10.

M Janina Sarol, Gibong Hong, Evan Guerra, and Halil Kilicoglu. 2024. Integrating deep learning architectures for enhanced biomedical relation extraction: a pipeline approach. *Database*, 2024:baae079.

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.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194.

Anthony Yazdani, Ihor Stepanov, and Douglas Teodoro. 2025. Gliner-biomed: A suite of efficient models for open biomedical named entity recognition.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880.

A. Appendix A. Oracle System Upper Bound Analysis

To quantify the upper bound of the system, we performed an oracle experiment in which gold test spans were passed directly to KRISSBERT, bypassing Stage 1 entirely. Table 10 compares this setting against our best end-to-end configuration.

Setting	Recall@1	Recall@5	Recall@10	MRR	mIoU
End-to-end (best)	72.65	78.46	78.99	0.752	0.3869
Oracle	85.43	94.89	96.12	0.896	0.5384

Table 10: Linking performance with perfect (gold) span detection compared to end-to-end.

With perfect span detection, KRISSBERT achieves a macro-IoU of 0.5384 and Recall@10 of 96.1% which is a substantial improvement over the end-to-end system. Moreover, the linker retrieves the correct concept for the majority of mentions within the top candidates. However, the gap between Recall@1 (85.4%) and Recall@5 (94.9%) indicates that the primary challenge lies in disambiguation among semantically similar candidates. This observation is consistent with the candidate ranking analysis in Table 9. The remaining approximately 4% of mentions for which the gold concept is not retrieved within the top 10 candidates represents an upper bound on retrieval performance.

Temporal Structure in Clinical Narratives in Portuguese: Insights from Cross-Document Annotation

Ana Luísa Fernandes[*‡§], **Purificação Silvano**[*‡§] **Luís Filipe Cunha**[*†]
[*] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)
[‡] Faculty of Arts and Humanities, University of Porto
[†] Faculty of Sciences, University of Porto
[§] Centre for Linguistics of the University of Porto (CLUP)
Porto, Portugal
{ana.l.fernandes, purificacao.silvano, luis.f.cunha}@inesctec.pt

Abstract

Medical reports, by documenting disease progression and patient responses to treatment, form continuous narratives in which each new document adds a chapter to the patient's clinical story. Constructing coherent patient timelines requires identifying temporal relations across multiple medical reports that compose a patient's clinical journey. However, cross-document temporal annotation remains an underexplored area, largely due to the methodological and conceptual challenges it entails. This study addresses these challenges by investigating the identification and characterization of cross-document temporal relations in Portuguese medical records. For this purpose, cross-document annotation was performed on different types of reports (Group Consultation Reports, Discharge Reports, and General Reports) from patients diagnosed with Acute Myeloid Leukemia and followed at IPO-Porto, Portugal. Annotation was carried out using the Med2Story scheme, specifically designed to capture both temporal and medical information. Our results indicate that, although cross-document annotation of temporal information is more demanding in terms of both the annotation scheme and the annotation process, it enables the construction of coherent chronological representations of patients' clinical journeys. Furthermore, the analysis reveals key characteristics of these clinical narratives, including the predominance of nominal events and the prevalence of simultaneity as the most frequent temporal relation type.

Keywords: clinical narratives, temporal annotation, cross-document annotation, Electronic Health Records

1. Introduction

Information extraction from Electronic Health Records (EHRs) remains a particularly challenging task. These records contain extensive amounts of unstructured narrative text, which introduces major challenges for data organization, curation, management, and effective reuse in both clinical and research settings. Natural Language Processing (NLP) methods play an essential role in automating the retrieval, processing, and extraction of relevant biomedical information from these records (Irrera et al., 2024).

High-quality NLP methods require manually annotated clinical data for their development. However, the development of high-quality annotated corpora remains especially demanding in under-resourced languages such as Portuguese. This limitation is particularly evident in the study of diseases with poor prognosis, such as Acute Myeloid Leukemia (AML). AML is a malignancy of the bone marrow stem cells, typically associated with an unfavorable outcome (Shimony et al., 2023).

In such cases, understanding a patient's clinical trajectory depends on reconstructing the temporal sequence of events documented across multiple medical reports. In this context, cross-document temporal analysis, aimed at generating chronologically coherent patient timelines, can yield significant contributions to temporal semantics and, simultaneously, to clinical NLP applications. This process involves identifying and labeling events, capturing their morphosyntactic and semantic features, and establishing temporal relations between events that occur across different documents. It is, however, a demanding and labor-intensive task (Zhao et al., 2023), requiring fine-grained interpretation and consistent annotation across heterogeneous clinical narratives.

Although several studies have addressed temporal annotation in clinical narratives, these efforts focus primarily on the temporal annotation of single documents rather than inter-document relations (Sun et al., 2013b; Galescu and Blaylock, 2012). To the best of our knowledge, only Raghavan et al. (2014) have proposed a model for cross-narrative temporal ordering of medical events, although they did not present a systematic annotation scheme for this purpose. Some efforts have been made in other domains, such as Minard et al. (2015), whose Cross-Document Event Ordering Task aimed to build coherent timelines from English news articles.

As a matter of fact, the cross-document annotated datasets of clinical narratives are scarce as the annotation is a complex and cognitively demanding process. Clinical events are often dis-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 300–312
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tributed across multiple documents produced at different times and by different healthcare professionals, leading to inconsistencies in terminology, narrative structure and level of detail. Moreover, clinical text differs substantially from non-clinical text due to its domain-specific language and the frequent use of abbreviations, which significantly increases the difficulty of processing data (Moharasan and Ho, 2019). Another challenge arises from the high heterogeneity of EHRs, as their content and writing style vary considerably across hospitals (Zhu et al., 2023), and even across different services and departments within the same institution. Consequently, reconstructing a patient's clinical timeline requires identifying and aligning events across heterogeneous sources while preserving temporal coherence. The present study aims to demonstrate how the challenges of cross-document annotation can be addressed in order to capture the temporal organization of medical reports. Specifically, it seeks to: (1) Develop and apply methodological strategies to overcome the main difficulties associated with cross-document temporal annotation in clinical texts; (2) Identify and describe cross-document temporal properties in medical records written in Portuguese for nine patients diagnosed with Acute Myeloid Leukemia at the Portuguese Institute of Oncology of Porto (IPO-Porto).

The paper is organized as follows. Section 2 presents the related work, contextualizing previous studies and resources relevant to this research. Section 3 describes the corpus selected for annotation, the annotation scheme (3.1), and the annotation methodology adopted (3.2). Finally, Section 4 reports the qualitative and quantitative results, focusing on cross-document annotation, but also presenting results related to intra-document analysis. Finally, Section 5 summarizes the main findings of the paper and outlines directions for future work.

2. Related Work

Medical reports, by documenting the progression of disease and the patient's response to treatment, constitute continuous narratives in which each new document adds a chapter to the patient's clinical story. To make informed decisions, physicians frequently consult previous reports to reconstruct the patient's clinical trajectory. Although NLP in medicine has received increasing attention, subfields such as anonymization and the extraction of clinical concepts (e.g., pathologies or medications) have received far more focus than the analysis of temporal structures (Sun et al., 2013b). The i2b2 project (Sun et al., 2013a,b) was pioneering in making available the first fully anonymized clinical corpus annotated with temporal information. Subsequently, Bethard et al. (2015); Bethard and Parker (2016); Bethard et al. (2017) organized the Clinical TempEval challenges, which used the THYME corpus (Styler IV et al., 2014). In parallel, other studies adapted TimeML (Pustejovsky et al., 2003) to the clinical domain, including Galescu and Blaylock (2012) and Savova et al. (2009), who proposed versions tailored to the specificities of medical narratives. The CLEF project (Roberts et al., 2007) annotated 167 clinical records, focusing on intra-phase relations and event–document creation time relations. More recently, the 2024 Shared Task on Chemotherapy Treatment Timeline Extraction introduced patient-level timelines as a novel challenge, involving tasks such as extracting relevant events, temporal expressions, and temporal relations from each document, and subsequently summarizing this information across documents (Yao et al., 2024).

In non-English-speaking contexts, important contributions include Campillos-Llanos et al. (2018), who developed the MERLOT corpus for French, and Campillos-Llanos et al. (2025), who introduced MEDSPANER, a medical and temporal named entity recognizer for Spanish, based on the TimeML framework and the HeidelTime system.

Despite the significant progress in intra-document temporal annotation, cross-document temporal annotation remains largely underexplored. Recent work has addressed event coreference resolution (Wu et al., 2020; Held et al., 2021; Eirew et al., 2021; Chen et al., 2025) and cross-document relation extraction (Yao et al., 2024; Jain et al., 2024). Gao et al. (2024a) proposed a cross-document coreference resolution model based on discourse information, integrating structural and semantic document representations through Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and lexical chains. Within the SemEval-2015 Task 4, Minard et al. (2015) focused on cross-document event coreference resolution and cross-document temporal relation extraction to construct timelines from English news articles. Caselli and Vossen (2017) aimed to extract and classify events relevant to narratives, drawing from news documents distributed over time and clustered around a single seminal event or topic. More recently, Gao et al. (2024b) developed a large-scale, domain-agnostic dataset for cross-document event extraction. While recent studies have expanded the scope of cross-document information extraction, several persistent challenges continue to hinder progress in temporal annotation across documents. The task imposes a substantial cognitive burden on annotators, who must read and interpret multiple texts, retain fine-grained details, and determine whether different mentions refer to the same

underlying event (Song et al., 2018). This process is particularly demanding when descriptions evolve over time or vary across sources, a common occurrence in clinical narratives, where similar procedures may be documented in distinct reports. Another obstacle arises from the absence of standardized guidelines for identifying and linking events across documents (Girardi et al., 2014), which undermines consistency and interoperability.

Building on the gaps identified in the literature, the present study aims to address the lack of research on cross-document temporal relations in clinical contexts. By doing so, this research contributes to advancing temporal annotation studies in the medical domain and to deepening the understanding of the temporal organization and narrative structure of Portuguese clinical records.

3. Dataset

The dataset to be annotated was extracted from a larger collection of 906 medical reports (MRs) written in European Portuguese, referring to 93 patients diagnosed with AML and followed at the IPO-Porto, Portugal. Access to these documents was authorized by the IPO-Porto Ethics Committee, framing the project within the institution's data management plan (Rb-Silva and Karimova, 2021).

The corpus included three types of reports: (1) Group Consultation Reports – MRs resulting from group consultations in which multiple specialists review the patient's clinical history as well as their clinical and analytical condition; (2) Discharge Reports – MRs documenting the patient's clinical evolution from hospital admission to medical discharge; (3) General Reports – MRs describing the patient's clinical trajectory up to the moment of writing, often requested to justify medical leave.

The examples presented in this paper correspond to adaptations of real sentences extracted from the MRs, which were previously pseudonymized to ensure that patient identification is impossible.

The creation of this dataset followed a set of predefined selection criteria to ensure consistency and representativeness. First, a maximum of 40 reports was set, deemed sufficient to ensure feasibility without compromising representativeness. Second, the inclusion of all three types of reports was ensured to guarantee typological diversity. Third, patients with more than two and fewer than five reports were selected, allowing for cross-document annotation while keeping homogeneity across patients. Finally, preference was given to documents with a narrative character, excluding those limited to laboratory results or complementary examinations.

Applying these criteria resulted in a subcorpus of 38 reports from 9 patients: 19 corresponding to Group Consultations, 16 to Discharges, and 3 to General Reports. The average document length was 157 words (standard deviation: 104). Within each patient's record, reports were organized and annotated in chronological order of document creation. A detailed distribution of report types per patient and their respective lengths is provided in the project's GitHub Repository[1].

3.1. Med2Story Annotation Scheme

The annotation scheme was designed to capture morphosyntactic, semantic, and domain-specific medical information, addressing the challenge of representing the temporal structure of a patient's clinical journey as described across multiple types of medical reports. The development of the annotation scheme was conducted in four main phases: (1) Foundational preparation, in which three existing annotation schemes were compared to identify their respective strengths and weaknesses, with the aim of replicating the strengths and avoiding the weaknesses; (2) Design and specification, which involved improving an existing temporal scheme and adding domain-specific medical tags; (3) Empirical validation, through a pilot annotation campaign involving multiple annotators; and (4) Consolidation and refinement, during which the scheme was further improved based on the validation results. The detailed methodology adopted in the design of the scheme is described in Fernandes et al. (2025b)

The annotation scheme is an extension of the *Text2Story* scheme (Silvano et al., 2021; Leal et al., 2022), developed on the basis of ISO 24617-1: 2012 . The scheme allows for the annotation of two types of structures Entity Structure and Link Structure, enabling the annotation of Events, Temporal Expressions, and Temporal Relations between Events, and between Events and Temporal Expressions.

Within the category of Events, a distinction is made between *General Events* and *Specialized Events*. General Events retain the original *Text2Story* attributes (namely *class*, *type*, *tense*, *aspect*, *polarity*, *vform*, *modality*, and *POS*) which are suitable for the linguistic description of clinical narratives. In contrast, Specialized Events incorporate domain-specific attributes tailored to the medical context, allowing for a more fine-grained and semantically enriched representation of clinical information. For example, in the sentence *"O hemograma mostrou leucocitose"* [*The blood count showed leukocytosis*], the events *"hemograma"* [*blood count*] and *"leucocitose"* [*leukocytosis*] would be annotated as Special-

[1] https://github.com/analuisacardosofernandes/Temporal-Structure-in-Clinical-Narratives

ized Events, whereas *"mostrou"* [*showed*] would be annotated as a General Event. Further details can be found in Fernandes et al. (2025a).

Figure 1 presents an overview of the annotation scheme. A detailed description of the scheme and

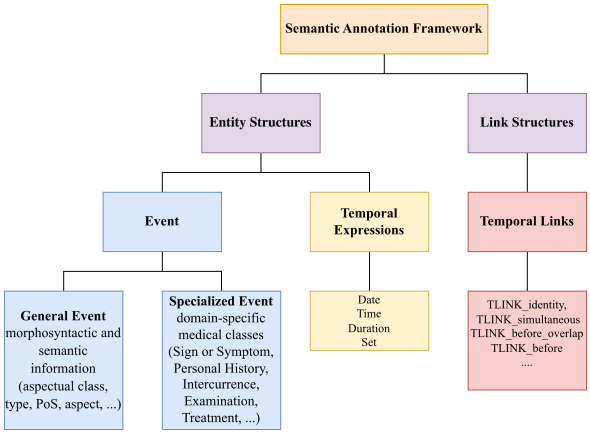


Figure 1: Annotation Scheme

its corresponding guidelines are available in the GitHub Repository.

3.2. Annotation Methodology

Annotation was performed by a single annotator with experience in semantic annotation and a background in Linguistics and Pharmaceutical Sciences. According to Roberts et al. (2009), the combination of linguistic and clinical knowledge in annotators tends to yield higher-quality annotations in this type of texts. The decision to employ a single annotator stemmed from the fact that morphosyntactic, semantic, and medical-domain annotation requires a highly specialized profile that is difficult to recruit. Furthermore, access to the original medical reports was restricted to a very limited number of researchers, preventing the inclusion of additional annotators.

To mitigate the limitations of having only one annotator, emphasis was placed on the clarification and validation of the guidelines, reviewed by annotators with experience in semantic annotation, and on the construction of a decision tree, validated by a hematology specialist. In cases of uncertainty regarding the classification of Specialized Events, the decision tree was consulted. Additionally, weekly meetings were held with a specialist in temporal semantics to discuss problematic cases, as well as regular meetings with the medical specialist to resolve domain-specific annotation questions. The validation process of the guidelines is described in detail in Fernandes et al. (2025b)

The annotation tool used was the INCEpTION platform (Klie et al., 2018). Within INCEpTION, nine documents (patient#1, patient#2, . . .) were created, each corresponding to the medical journey of an individual patient. Each document contains all the clinical reports associated with that patient's case. To improve clarity and facilitate the identification of report boundaries, each report within a document was separated by a blank line.

Annotation was carried out in two stages: first, events and temporal expressions, along with their attributes, were annotated across all reports, followed by the annotation of TLINKs. This approach follows Campillos-Llanos et al. (2018), who found that higher inter-annotator agreement (IAA) and overall annotation quality were achieved when the process was conducted in two steps: first annotating entities and their attributes, and then proceeding with relation annotation.

The reports of each patient were annotated in chronological order of their creation. Cross-document annotation followed a systematic procedure, as illustrated in the workflow presented in Figure 2. For instance, if the first report for a patient is a group consultation with a DOCTIME of *17.09.2023* and the next report is a Discharge Report with a DischargeTime of *28.01.2024*, this date must be linked to the DOCTIME of the first report by *TLINK_after*.

Additionally, when a Group Consultation or General Report was written during a hospitalization period, the DOCTIME of that report was linked to both the AdmissionTime and the DischargeTime recorded in the corresponding Discharge Report. For example, if the first report is a discharge summary with an AdmissionTime of *03.06.2022* and a DischargeTime of *06.08.2022*, and the second report is a group consultation with a DOCTIME of *05.06.2022*, this DOCTIME (*05.06.2022*) must be linked to the AdmissionTime (*03.06.2022*) by *TLINK_after* and to the DischargeTime (*06.08.2022*) by *TLINK_before*.

Whenever possible, a TLINK was also established between the first event of a given report and the last event of the preceding one. Events within a report were only linked to those of the previous report through *TLINK_identity*, except in cases where an event mentioned but not temporally located in the previous report could be temporally anchored by an event in the subsequent one.

All decisions were documented in the annotation guidelines, available in the GitHub Repository.

4. Results

With respect to the corpus[2] as a whole, a total of 1.859 markables were identified, of which 861

[2]Due to privacy and confidentiality constraints, the dataset cannot yet be shared publicly while ethical approvals and patient consent are being obtained.

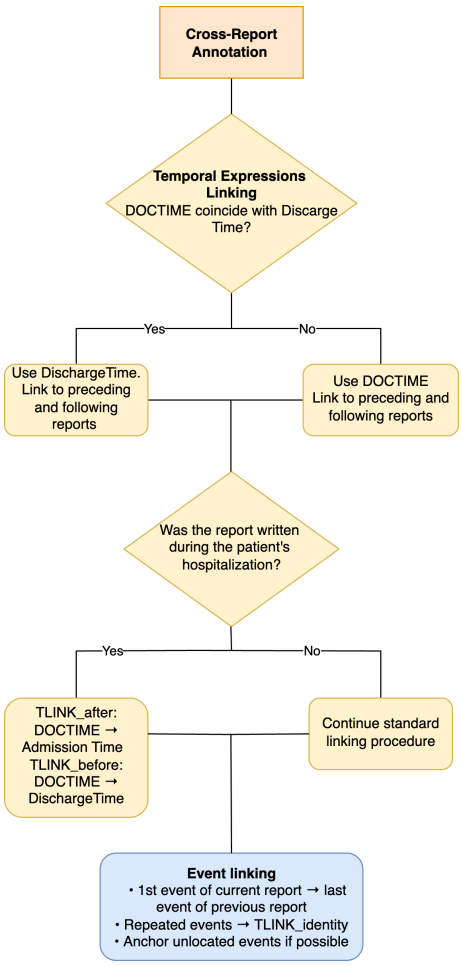


Figure 2: Cross-document annotation rules

(46%) correspond to Specialized Events, 728 (39%) to General Events, and 270 (15%) to Temporal Expressions. The main results are presented here, with detailed annotation results available in the GitHub Repository.

4.1. Events and Time Expressions

Among General Events, 50% (362) are nouns, 46% (335) verbs, and 4% (31) adjectives. Specialized Events reveal an even more marked predominance of nouns, accounting for 95.05% of all events (807 instances). Nominal expressions are used to refer to diagnoses and disease characterizations, symptoms, examinations and their results, procedures, and treatments. Adjectives constitute the second most frequent category in Specialized Events, representing 4.47% (38 instances). They serve mainly to qualify health states or clinical results, and also to express the modality through terms such as "provável" [*probable*]. Other categories occur only marginally: pronouns represent 1.40% (12 instances); verbs represent 0,23% (2 instances); and other categories, including adverbs (e.g., “o paciente está bem” [*the patient is well*]), account for 0.23%. The prevalence of eventive nouns appears to be characteristic of this textual genre, contrasting with journalistic narratives, where they occur less frequently (Silvano et al., 2024).

Table 1: Distribution of General Event Class

Class	Frequency	Percentage
Occurrence	166	22.80%
State	145	19.92%
I_Action	31	4.26%
Reporting	13	1.78%
None	373	51.24%

Regarding polarity in General Events, 96.98% of the cases (706 instances) display positive polarity, and only 3.02% (22 instances) negative polarity, mostly restricted to expressions such as "sem queixas" [*no complaints*]. This pattern aligns with the nature of medical reports, which predominantly describe clinical facts through affirmative statements, recording what is present rather than what is absent. Regarding Specialized Events, most events display positive polarity (836 instances), with only 25 exhibiting negation. Certain Specialized Event Classes show a comparatively higher rate of negation, notably Examination Result (e.g., “sem isolamentos bacteriológicos” [*no bacteriological isolations*]), followed by Personal History (e.g., “sem antecedentes pessoais de relevo” [*no relevant personal history*]), and Treatment Response (e.g., “sem recuperação hematológica” [*no hematological recovery*]).

General Events are mainly dynamic (Table 1, reflecting the narrative dimension of medical reports, where events advance the temporal progression of discourse (Caenepeel, 1989) (e.g., "iniciou protocolo de indução" [*he initiated induction protocol*]). States, though frequent, serve descriptive, characterizing the patient’s condition, providing personal data, or reporting examination results (e.g., "quadro recente caracterizado por astenia" [*recent condition characterised by astenia*]).

As shown in Table 1, 51.24% of the events were not assigned aspectual features (label *none*), since these were applied only to verbal events, and not to nominal ones, given that there is currently no established approach in the literature for the aspectual annotation of nominal events.

Reporting verbs (e.g., "dizer" [*to say*], "afirmar" [*to state*]) occur infrequently, suggesting that physicians seldom reproduce patients’ statements through indirect speech. Patient information is instead presented as factual data, devoid of enunciative responsibility. The same applies to symptoms, which, although subjective phenomena reported by the patient to the doctor, are expressed in an objective manner (e.g., "Instalou-se um quadro de sudorese nocturna, astenia e mialgias" [*a condition of night sweats, asthenia, and myalgia developed*]).

Table 2: Distribution of General Event Type

Type	Frequency	Percentage
Transition	165	22.66%
State	145	19.92%
Process	45	6.18%
None	373	51.24%

This absence of reported speech confirms observations made by Lysanets et al. (2017), who identify it as a distinctive feature of medical discourse. In the corpus, reporting events were limited to verbs such as "realçar" [*to emphasise*] and "referir" [*to note*], differing markedly from journalistic texts, which employ a range of enunciative strategies and direct speech. Silvano et al. (2024) report an average of 3.50 reporting events per news item.

No *Perception* events or *I_State* events (intensional states such as "acreditar" [*to believe*]) were found, reflecting the high degree of objectivity in medical descriptions. *I_Action* events occurred only in the construction "decide-se propor" [*it was decided to propose*]. Savova et al. (2009) also noted that the less common event classes in clinical narratives were *Perception* and *Reporting*.

Regarding General Event type (Table 2), *Transition* was the most frequent class, followed by *State*.

Concerning *VForm*, annotated only for non-finite verbs, 76 instances of the participle , 46 of the infinitive, and 20 of the gerund were identified. Finite forms were more common (586 occurrences). Only two instances of the future and one of the subjunctive were found. In terms of modality, "poder" [*may*] appeared six times and "dever" [*must*] once. The low frequency of modal forms reflects the communicative purpose of medical reports, which favor factual information over hypothetical or evaluative content.

Most events were expressed in the past tense, in line with the retrospective nature of clinical narratives. Lysanets et al. (2017) similarly note that the predominance of the past tense in medical case reports stems from their narrative purpose: to recount past events. Sixty-four events occurred in the present tense, typically describing the patient's current status, and only two in the future, both expressing epistemic modality rather than temporal reference (e.g., "será de realçar uma hiperbilirrubinemia" [*it should be noted that there is hyperbilirubinemia*]).

Regarding the Specialized Events, the most predominant was Treatment (Table 3).

The centrality of treatment reflects the very function of medical reports: all the information they contain ultimately serves to guide the clinician's therapeutic decisions. Group Consultation Reports systematically conclude with a treatment proposal; Discharge Reports necessarily include treatment-related references, since hospitalization itself is

Table 3: Distribution of Specialized Event Classes

Specialized Event Class	Frequency	Percentage
Treatment	167	19.40%
Examination Result	133	15.45%
Personal History	103	11.96%
Examination	99	11.50%
Sign or Symptom	83	9.64%
Characterization of the Disease	82	9.53%
Principal Diagnosis	66	7.66%
Intercurrence	56	6.50%
Treatment Response	51	5.92%
Medical Procedure	14	1.63%
Drug Administration Route	7	0.81%

linked to therapy (either for chemotherapy or for an intercurrent event arising from it); and General Reports incorporate treatment within the retrospective overview of the patient's medical course. In contrast, Drug Administration Route appears least frequently, with only seven occurrences limited to the specifications intravenous, intrathecal, and oral. This scarcity suggests that, although treatment constitutes the thematic axis around which the reports are organized, emphasis is placed on decision-making and outcomes rather than on the technical details of drug administration.

Finally, several abbreviations were identified in the corpus and subsequently clarified during the annotation process in an open-text field, in order to ensure terminological transparency (e.g., LMA ("Leucemia Mieloide Aguda" [*Acute Myeloid Leukemia*]). Most of these abbreviations correspond to disease names.

With regard to Temporal Expressions, the most frequent category was Date (Table 4). Expressions of this type mainly refer to dates of hospitalization, discharge, consultations, interventions, and to the beginning and end of treatments.

Table 4: Distribution of Temporal Expressions

Temporal Expressions	Frequency	Percentage
Date	76.29%	206
Duration	17.04%	46
Time	5.19%	14
Set	1.48%	4

The next most common category is Duration (e.g., "apiraxia sustentada durante 7 dias" [*apyrexia for 7 days*]), which essentially corresponds to the length of symptoms or treatments. This is followed by Time, and, finally, Set. The latter is generally associated with the frequency of medication administration or the periodic performance of examinations.

As for Temporal Function, most events were not assigned any annotation for this attribute (213 occurrences, 78.75%). Among the annotated cases, the most frequent label was DOCTIME (which refers to the date when the report was created) (6.65%, 18 occurrences), followed by Discharge-

Time and AdmissionTime (both 5.91%, 16 occurrences each), and, finally, AnchorTime (2.59%, 7 occurrences). This distribution suggests that only a small proportion of temporal expressions are directly associated with specific reference points in the clinical cycle (admission, discharge, or the document's reference time), while the majority are used descriptively, serving only to temporally situate events rather than perform an explicit temporal function.

4.2. TLINKs

TLINK_simultaneous was the most frequent (Table 5), reflecting the prevalence of events that occur concurrently or whose temporal boundaries cannot be meaningfully distinguished in context. This relation was also applied by default when no information enabled a more specific temporal localization. For example, in the sentence *"O hemograma mostrou leucocitose"* [*The blood count showed leukocytosis*], the events *"hemograma"* [*blood count*], *"mostrou"* [*showed*] and *"leucocitose"* [*leukocytosis*] would be linked with *TLINK_simultaneous* since the information regarding the beginning of "leucocitose is not available.

The high frequency of *TLINK_simultaneous* aligns with the high frequency of states, which naturally favor simultaneity links; however, dynamic events also frequently exhibit this relation, suggesting that medical reports, although they have a narrative character, often emphasize descriptive rather than sequential information.

Table 5: Distribution of TLINKs

TLINKs	Frequency	Percentage
TLINK_simultaneous	733	34%
TLINK_identity	345	16%
TLINK_after	337	15.63%
TLINK_isIncluded	286	13.27%
TLINK_includes	115	5.33%
TLINK_before	155	7.19%
TLINK_before_overlap	30	1.39%
TLINK_after_overlap	36	1.67%
TLINK_begins	40	1.85%
TLINK_begunBy	37	1.72%
TLINK_ends	12	0.56%
TLINK_endedBy	5	0.23%
TLINK_during	25	1.16%

TLINK_identity was the next most frequent, marking situations or temporal expressions that refer to the same event or the same time expression. Its recurrence is partly due to annotation rules based on the ISO standard, notably those governing light verb constructions (e.g., "fez tratamento" [*underwent treatment*]), where the verb and the nominal component are linked by *TLINK_identity* to represent a single event.

TLINK_after expresses chronological succession, and its relatively high frequency reflects the abundance of transition-type events, which inherently promote temporal sequencing.

Relations of inclusion also occurred frequently, capturing nested temporal relations, such as events occurring within a broader interval ("Fez quimioterapia de indução. Como intercorrências teve 2 síndromes febris" [*She underwent induction chemotherapy. As intercurrences, she experienced two febrile syndromes*]). Less frequent, though still relevant, were *TLINK_before* and overlap variations (*TLINK_before_overlap* and *TLINK_after_overlap*), typically associated with modifiers like "recentemente" [*recently*], which indicate partial temporal overlap. Finally, relations specifying explicit boundaries — *TLINK_begins*, *TLINK_begunBy*, *TLINK_ends*, *TLINK_endedBy*, and *TLINK_during* — most often occurred in medication contexts, such as the beginning and end of drug administration.

4.3. Cross-document Temporal Analysis Across Medical Reports

Table 6 of Appendix A presents the cross-document relations that were established. The predominant temporal relation between successive reports of the same patient is, as expected, *TLINK_identity*, since personal history, diagnosis, and treatment are repeated throughout the reports. The events most frequently associated with this relation belong to the Specialized Event and are predominantly nominal. The most common classes linked by this TLINK include Treatment, Principal Diagnosis, Characterization of the Disease, and Personal History. These results are consistent with the nature of clinical reports: Treatment is reiterated throughout the document sequence, being initially mentioned in the Group Consultation Report (when the patient is first admitted to the service), resumed during hospitalization for treatment, and referred to again in subsequent, more general reports. Similarly, the Principal Diagnosis, Acute Myeloid Leukemia, and the Characterization of the Disease tend to be recurrently linked across reports, as the patient's entire clinical history revolves around this main diagnosis. Personal History elements are also frequently repeated. The second most frequent relations are *TLINK_after* and *TLINK_before*, which occur predominantly between temporal expressions of the Date type, particularly between DOCTIMEs, between DischargeTime and DOCTIME, or among different DischargeTimes. These relations thus establish the chronological progression between reports. *TLINK_after* also appears linking General Events across different documents. Most of these events are nominal and therefore lack

explicit aspectual specification. In cases where the General Event is verbal, *TLINK_identity* relations are mainly observed between two states, and *TLINK_before* between two occurrences of the transition type. Less frequent relations also emerge, such as *TLINK_is_included*. This relation was observed, for instance, when the last event in the first report was "chemotherapy" and the first event in the subsequent report was "discharge". In this case, the "discharge" event occurred during the "chemotherapy" period and is therefore considered to be included within it.

In what follows, an example of the cross-document annotation is provided, corresponding to the record of Patient 2. The record of Patient 2 comprised one *Discharge Report*, three *Group Consultation Reports*, and one *General Report*, amounting to a total of 201 annotated relations. Regarding cross-document annotation, the following *TLINKs* were established: (1) **Six *TLINK_after*** — two between *Group Consultation Report #1* and *Discharge Report #1*; one between *Group Consultation Report #2* and *Group Consultation Report #1*; one between *Group Consultation Report #3* and *Group Consultation Report #2*; and one between *General Report #1* and *Group Consultation Report #2*; (2) **Twenty-six *TLINK_identity*** — four between *Group Consultation Report #1* and *Discharge Report #1*; seven between *Group Consultation Report #2* and *Group Consultation Report #1*; six between *Group Consultation Report #3* and *Group Consultation Report #2*; and nine between *General Report #1* and *Group Consultation Report #3*; (3) **Four *TLINK_before*** — one between *Group Consultation Report #1* and *Discharge Report #1*; one between *Group Consultation Report #2* and *Group Consultation Report #1*; one between *Group Consultation Report #3* and *Group Consultation Report #2*; and one between *General Report #1* and *Group Consultation Report #3*.

4.3.1. Challenges in Cross-document Annotation

Cross-document annotation presents a range of challenges that stem from cognitive workload and the inherent complexity of certain relation types. In the present work, several key difficulties were encountered. First, the task imposes a high cognitive load on the annotator. To establish TLINKs, annotators must read, understand, and retain information distributed across multiple clinical reports. This process requires not only sustained attention and memory, but also a degree of domain knowledge to recognize whether differently phrased descriptions refer to the same underlying clinical event and to determine the appropriate temporal placement of those events. The need to mentally compare and integrate details across documents makes the annotation process particularly demanding.

Ambiguity in decision-making is another recurring challenge. Determining whether two expressions refer to the same real-world event is not always straightforward, especially when descriptions differ in specificity or narrative focus. This difficulty is compounded by the distribution of relevant information across documents: events are often mentioned only in fragments. For instance, one report may state that a patient has been "followed in this service for four years", while another specifies "followed in this service since April 2014", providing a clearer temporal anchor. As a result, cross-document annotation frequently involves linking partial and complementary pieces of information.

A further complication arises from evolving and sometimes conflicting clinical records. The same event may be described differently over time or by different healthcare professionals, reflecting updates, corrections, or documentation inconsistencies. In some cases, the same event was assigned different dates across reports — for example, a cesarean section listed as occurring in September 2014 in one report and in October 2014 in another. Such discrepancies challenge the establishment of *TLINK_identity*, as annotators must decide whether new or corrected details refer to the same event or to a distinct one. Finally, as annotation progresses, documents in INCEpTION become increasingly dense with marked entities and relations. This density can reduce the usability of annotation tools, making navigation more difficult and increasing the likelihood of overlooking inconsistencies.

5. Final Remarks

In this study, we aimed to identify and describe cross-document temporal properties in medical records. The findings show that, although cross-document temporal annotation is more demanding in terms of both the annotation scheme and the annotation process, it enables the construction of coherent chronological representations of patients' clinical journeys. The analysis also highlights key characteristics of these narratives, such as the predominance of nominal event expressions and the frequent use of simultaneity as the primary temporal relation type. Future work will include extending the annotation scheme to incorporate participant and discourse-relation layers, exploring automatic timeline generation from annotated data, and investigating automatic pre-annotation methods to support and accelerate the annotation process.

6. Limitations

This study presents some limitations that should be acknowledged. First, the annotation was carried

out by a single annotator. Although measures were implemented to enhance the reliability and feasibility of the annotated data, as described in this article, the absence of inter-annotator agreement metrics may limit the ability to quantitatively assess annotation consistency and replicability.

Second, due to privacy and confidentiality constraints associated with clinical data, the annotated dataset cannot be publicly released. Although de-identification procedures were applied, the sensitive nature of the records prevents open access. This limitation reduces the reproducibility of the study and restricts external validation or reuse of the resource by the research community. In future work, this constraint will be addressed through the generation of synthetic clinical reports derived from the characterization of the original documents. Defining the structural and linguistic properties of this type of text, as undertaken in the present study, is therefore a necessary step toward enabling the creation of representative and shareable synthetic datasets.

Finally, the clinical reports analyzed originate from a single healthcare institution. As documentation practices, narrative styles, and temporal structuring in clinical reporting can vary across institutions and regions, the findings may not generalize to broader contexts. Future work should consider expanding the dataset to include documents from multiple sources in order to improve representativeness and enhance the general applicability of the annotation framework and results.

7. Acknowledgements

This work was supported by national funds through the Fundação para a Ciência e a Tecnologia (FCT), under PhD grant 2025.00440.BD. The authors also acknowledge support from the StorySense project (DOI: 10.54499/2022.09312.PTDC).

8. Bibliographical References

J. Vasco Barreto and Paulo Paiva. 2008. O registo clínico orientado por problemas = the problem oriented clinical record. *Medicina Interna: Revista da Sociedade Portuguesa de Medicina Interna*, 15(3):201–206.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, Denver, Colorado. Association for Computational Linguistics.

Steven Bethard and Jonathan Parker. 2016. A semantically compositional annotation scheme for time normalization. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3779–3786, Portorož, Slovenia. European Language Resources Association (ELRA).

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

Peter Bramsen, Pranav Deshpande, Young-Kyoon Lee, and Regina Barzilay. 2006. Finding temporal order in discharge summaries. In *AMIA Annual Symposium Proceedings*, volume 2006, pages 81–85. American Medical Informatics Association.

Mimo Caenepeel. 1989. *Aspect, temporal ordering and perspective in narrative fiction*. Phd thesis in linguistics, University of Edinburgh, Edinburgh.

Leonardo Campillos-Llanos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. 2018. A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, 52(2):571–601. Online publication.

Leonardo Campillos-Llanos, Ana Valverde-Mateos, and Adrián Capllonch-Carrión. 2025. Hybrid natural language processing tool for semantic annotation of medical texts in spanish. *BMC Bioinformatics*, 26(1):7.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Xinyu Chen, Peifeng Li, and Qiaoming Zhu. 2025. Improving cross-document event coreference resolution by discourse coherence and structure. *Information Processing & Management*, 62(4):104085. [Online].

Alon Eirew, Arie Cattan, and Ido Dagan. 2021. WEC: Deriving a large-scale cross-document event coreference dataset from Wikipedia. In *Proceedings of the 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2498–2510, Online. Association for Computational Linguistics.

Ana Luísa Fernandes, Purificação Silvano, António Leal, Nuno Guimarães, Evelin Amorim, and Alípio Jorge. 2025a. Can ISO 24617-1 go clinical? extending a general-domain scheme to medical narratives. In *Proceedings of the 21st Joint ACL–ISO Workshop on Interoperable Semantic Annotation (ISA-21)*, pages 41–52, Düsseldorf, Germany. Association for Computational Linguistics.

Ana Luísa Fernandes, Purificação Silvano, António Leal, Nuno Guimarães, Rita Rb-Silva, Luís Filipe Cunha, and Alípio Jorge. 2025b. The incremental process of building an annotation scheme for clinical narratives in Portuguese: the contribution of human variation analysis. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 332–343, Vienna, Austria. Association for Computational Linguistics.

Lucian Galescu and Nathan Blaylock. 2012. A corpus of clinical narratives annotated with temporal information. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI 2012)*, pages 715–720, Miami, Florida. Association for Computing Machinery (ACM).

Qiang Gao, Bobo Li, Zixiang Meng, Yunlong Li, Jun Zhou, Fei Li, Chong Teng, and Donghong Ji. 2024a. Enhancing cross-document event coreference resolution by discourse structure and semantic information. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5907–5921, Torino, Italia. ELRA and ICCL.

Qiang Gao, Zixiang Meng, Bobo Li, Jun Zhou, Fei Li, Chong Teng, and Donghong Ji. 2024b. Harvesting events from multiple sources: Towards a cross-document event extraction paradigm. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1913–1927, Bangkok, Thailand. Association for Computational Linguistics.

Christian Girardi, Manuela Speranza, Rachele Sprugnoli, and Sara Tonelli. 2014. CROMER: a tool for cross-document event and entity coreference. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3204–3208, Reykjavik, Iceland. European Language Resources Association (ELRA).

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1406–1417, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ornella Irrera, Stefano Marchesin, and Gianmaria Silvello. 2024. Metatron: advancing biomedical annotation empowering relation annotation and collaboration. *BMC Bioinformatics*, 25:112.

Ann K. Irvine, Stephanie W. Haas, and Tessa Sullivan. 2008. Tn-ties: A system for extracting temporal information from emergency department triage notes. In *AMIA Annual Symposium Proceedings*, volume 2008, pages 328–332. American Medical Informatics Association.

ISO. 2012. *ISO 24617-1: 2012, Language resource management – Semantic annotation framework (SemAF) – Part 1: Time and events (SemAF-Time, ISO-TimeML)*. The International Organization for Standardization, Geneva. Editorial committee: James Pustejovsky (chair), Project leader and convenor: Kiyong Lee.

Monika Jain, Raghava Mutharaju, Kuldeep Singh, and Ramakanth Kavuluru. 2024. Knowledge-driven cross-document relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3787–3797, Bangkok, Thailand. Association for Computational Linguistics.

J.-C. Klie, M. Bugert, B. Boullosa, R. Eckart de Castilho, and I. 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. Association for Computational Linguistics.

António Leal, Purificação Silvano, Evelin Amorim, Inês Cantante, Fátima Silva, Alípio Mário Jorge, and Ricardo Campos. 2022. The place of iso-space in text2story multilayer annotation scheme. In *Proceedings of the 18th Joint ACL–ISO Workshop on Interoperable Semantic Annotation within LREC 2022*, pages 61–70, Marseille, France. European Language Resources Association.

Yu-Kai Lin, Hsinchun Chen, and Randall A. Brown. 2013. Medtime: A temporal information extraction system for clinical narratives. *Journal of Biomedical Informatics*, 46(Suppl.):S20–S28.

Yuliia Lysanets, Halyna Morokhovets, and Olena Bieliaieva. 2017. Stylistic features of case reports as a genre of medical discourse. *Journal of Medical Case Reports*, 11(83). Online.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional description of text organization. *Text*, 8:243–291.

Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, and Rubén Urizar. 2015. SemEval-2015 task 4: TimeLine: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 778–786, Denver, Colorado. Association for Computational Linguistics.

Gandhimathi Moharasan and Tu Bao Ho. 2019. Extraction of temporal information from clinical narratives. *Journal of Healthcare Informatics Research*, 3(2):220–244.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*, pages 28–34, Stanford, CA. AAAI Press.

Preethi Raghavan, Eric Fosler-Lussier, Noémie Elhadad, and Albert M. Lai. 2014. Cross-narrative temporal ordering of medical events. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 998–1008, Baltimore, MD. Association for Computational Linguistics.

Rita Rb-Silva and Yulia Karimova. 2021. amile: Application of text mining to clinical reports of patients with acute myeloid leukemia. Zenodo. Zenodo, January 28, 2021.

Angus Roberts, Richard Gaizauskas, Mark Hepple, George Demetriou, Yulan Guo, Andrea Setzer, and Bill Wheeldin. 2007. The clef corpus: semantic annotation of clinical text. In *AMIA Annual Symposium Proceedings*, pages 625–629.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, 42(5):950–966.

Guergana Savova, Steven Bethard, William Styler IV, James Martin, Martha Palmer, Julie Masanz, and Wayne Ward. 2009. Towards temporal relation discovery from the clinical narrative. In *AMIA Annual Symposium Proceedings*, pages 568–572.

Shira Shimony, Michael Stahl, and Richard M. Stone. 2023. Acute myeloid leukemia: 2023 update on diagnosis, risk-stratification, and management. *American Journal of Hematology*, 98(3):502–526. Epub 2023 Jan 13.

Purificação Silvano, António Leal, Fátima Silva, Inês Cantante, Fatima Oliveira, and Alípio Mario Jorge. 2021. Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus. In *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 1–13, Groningen, The Netherlands (online). Association for Computational Linguistics.

Purificação Silvano, Emanuel Amorim, António Leal, Inês Cantante, Adriana Jorge, Ricardo Campos, and Nanjiang Yu. 2024. Untangling a web of temporal relations in news articles. In *Proceedings of the 7th International Workshop on Narrative Extraction from Texts (Text2Story'24)*, pages 77–92, Glasgow. CEUR Workshop Proceedings (CEUR-WS.org).

Zhiyi Song, Ann Bies, Justin Mott, Xuansong Li, Stephanie Strassel, and Christopher Caruso. 2018. Cross-document, cross-language event coreference annotation using event hoppers. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013a. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46(Suppl 0):S5–S12. Epub 2013 Jul 19.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013b. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813. Epub 2013 Apr 5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lawrence L. Weed. 1968. Medical records that guide and teach. *The New England Journal of Medicine*, 278:593–600, 652–657. Online.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli Goldner, and Guergana Savova. 2024. Overview of the 2024 shared task on chemotherapy treatment timeline extraction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 557–569, Mexico City, Mexico. Association for Computational Linguistics.

Jin Zhao, Nianwen Xue, and Bonan Min. 2023. Cross-document event coreference resolution: Instruct humans or instruct GPT? In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 561–574, Singapore. Association for Computational Linguistics.

Enwei Zhu, Qilin Sheng, Huanwan Yang, Yiyang Liu, Ting Cai, and Jinpeng Li. 2023. A unified framework of medical information annotation and extraction for chinese clinical text. *Artificial Intelligence in Medicine*, 142:1–12.

A. Cross-document relations

Patients / Reports	MR #2 → MR #1	MR #3 → MR #2	MR #4 → MR #3	MR #5 → MR #4	MR #5 → MR #3	MR #4 → MR #1	MR #4 → MR #2	MR #3 → MR #1
Patient 1	After–2 Identity–5							
Patient 2	After–2 Before–1 Identity–4	After–1 Before–1 Identity–7	After–1 Before–1 Identity–6	After–1 Before–1 Identity–4	Identity–5			After–1
Patient 3	After–2 Identity–6	After–1 Identity–1	After–2 Identity–6	After–2 Identity–1				
Patient 4	After–2 Identity–1	After–1 Before–1 Identity–5						
Patient 5	After–1 Before–1	After–2 Begun_By–1						Identity–2
Patient 6	After–1 Before–1 Identity–4	After–3 Identity–11 Simultaneous–1	After–2 Identity–17	After–2 Begins–1 Identity–17	Identity–1			Includes–1 Identity–1
Patient 7	After–2	After–1 Before–1 Identity–2	After–1 Is_Included–2 Identity–2	After–2 Identity–3		Identity–2	Identity–1	
Patient 8	After–1 Is_Included–1 Identity–10	After–2 Identity–12	After–2 Identity–8	After–2 Identity–7				Identity–1
Patient 9	After–2 Identity–7	After–2 Identity–6	After–2 Identity–1	After–2 Identity–3		Identity–2		Identity–18

Table 6: Cross-document relations across MRs

MOSAIC: A Multilingual, Taxonomy-Agnostic, and Computationally Efficient Approach for Radiological Report Classification in Low-Resource Settings

Alice Schiavone[1,2], **Marco Fraccaro**[3], **Lea Marie Pehrson**[1,4,5], **Silvia Ingala**[2,4], **Rasmus Bonnevie**[3], **Michael Bachmann Nielsen**[1,4,5], **Vincent Beliveau**[7], **Melanie Ganz**[1,2], **Desmond Elliott**[1]

[1]Department of Computer Science, University of Copenhagen

[2]Neurobiology Research Unit, Copenhagen University Hospital

[3]Unumed ApS, [4]Department of Diagnostic Radiology, Copenhagen University Hospital

[5]Department of Clinical Medicine, University of Copenhagen

[5]Cerebriu A/S, [7]Institute for Human Genetics, Medical University of Innsbruck

Abstract

Radiology reports contain rich clinical information that can be used to train imaging models without relying on costly manual annotation. However, existing approaches face critical limitations: rule-based methods struggle with linguistic variability, supervised models require large annotated datasets, and recent LLM-based systems depend on closed-source or resource-intensive models that are unsuitable for clinical use. Moreover, current solutions are largely restricted to English and single-modality, single-taxonomy datasets. We introduce MOSAIC, a multilingual, taxonomy-agnostic, and computationally efficient approach for radiological report classification. Built on a compact open-access language model (MedGemma-4B), MOSAIC supports both zero-/few-shot prompting and lightweight fine-tuning, enabling deployment on consumer-grade GPUs. We evaluate MOSAIC across seven datasets in English, Spanish, French, and Danish, spanning multiple imaging modalities and label taxonomies. The model achieves a mean macro F1 score of 88 across five chest X-ray datasets, approaching or exceeding expert-level performance, while requiring only 24 GB of GPU memory. With data augmentation, as few as 80 annotated samples are sufficient to reach a weighted F1 score of 82 on Danish reports, enabling large-scale cohort classification with minimal human effort. Code and models are open-source, offering a practical alternative to large or proprietary LLMs in clinical settings.

Keywords: radiology reports classification, multilingual NLP, low-resource learning

1. Introduction

Deep learning methods have been extensively explored for AI-assisted medical imaging analysis. However, their effectiveness depends on large volumes of annotated data. Such annotations must be provided by expert radiologists, whose primary focus remains clinical care, limiting the availability of high-quality labeled datasets. A promising solution to the annotation bottleneck is to extract relevant information directly from radiology reports, which are routinely produced during imaging procedures to document abnormalities associated with clinical findings (Reichenpfader et al., 2024). This information can later be used to train medical imaging classifiers or perform retrospective clinical studies (Zhou et al., 2014). Classic methods to automate finding extraction from English radiology reports include rule-based methods and BERT-based classifiers, that can prove effective in providing annotations close to those of radiologists (Irvin et al., 2019; Smit et al., 2020). However, rule sets need to be hand-crafted and are still limited by syntactic variability, and deep learning methods need a large amount of annotations from expert clinicians (Yang et al., 2023). Adapting these methods to a new label taxonomy or language necessitates either retraining the model, partially or entirely, or developing new rule sets from scratch. Re-training models to adapt to a new language or taxonomy for the same underlying task is an inefficient use of computational resources and is not sustainable.

In contrast, recent advances in natural language processing have led to the emergence of large language models (LLMs), which follow user instructions through natural language prompting. These models enable zero-shot or few-shot classification (i.e., with no or very few labeled examples), eliminating the need for manual annotation or rule engineering, and offering greater adaptability across tasks, languages, and taxonomies (Gu et al., 2024; Dorfner et al., 2024). However, the use of LLMs in many clinical settings often depends on large and/or closed-source models that cannot be deployed locally, posing significant challenges for projects involving sensitive patient data due to privacy concerns and the need for high-end computational resources. In these scenarios, compact models that can operate on consumer-grade hardware are the most practical solution.

While several studies have explored this task, they are typically restricted to a single dataset, limited language coverage (often one or two languages), and limited to a single medical imaging

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 313–323
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

modality, such as X-rays or magnetic resonance imaging (MRI), each having distinct clinical contexts and reporting conventions (Reis et al., 2022; Nguyen et al., 2022; Mottin et al., 2023; Wollek et al., 2024; Olivato et al., 2024; Matsuo et al., 2024; Collado-Montañez et al., 2025; Mergen et al., 2025; Al Mohamad et al., 2025). Even when datasets share the same imaging modality, differences in research focus lead to variability in labeled findings. For example, a chest X-ray may be used to study either the heart or the lungs, resulting in highly diverse label sets across studies. While LLMs offer greater flexibility than traditional deep learning methods, they still face important limitations. Most models are primarily trained on English, as it dominates the available web-crawled data. As a result, despite their strong performance on many tasks, LLMs underperform compared to BERT-based approaches when applied to other languages, e.g. Danish MRI reports (Beliveau et al., 2024), or Japanese pancreatic cancer reports (Suzuki et al., 2024).

To identify a viable approach to radiological report classification, we investigate whether generative language models can effectively perform this task in multiple languages and taxonomies, particularly in low-resource settings with limited computational capacity, scarce annotated data, and the need for localized model deployment. We look at language and task competency, evaluating performance across diverse linguistic settings and label taxonomies, and analyzing trade-offs between model size, accuracy, and adaptability.

We propose MOSAIC, an efficient and flexible LLM-based method for radiological report classification with the following key properties:

- **Multilingual**: Trained and evaluated across multiple languages, including English, Spanish, French, and Danish.
- **Optimized for Small-scale**: Designed to be trained and tested on consumer-grade GPUs, enabling local model development and adaptation while preserving patient privacy by avoiding reliance on external services.
- **Adaptable**: Robust to variations in label taxonomies across diverse datasets and medical imaging modalities.
- **Annotation-efficient**: With data augmentation, the model can match full-dataset performance using as few as 80 annotated samples.
- **Open-source**: We release the code and models in two sizes (4B and 12B)[1], while most datasets are accessible through their original providers.

[1]Available upon paper acceptance

We invite researchers to evaluate our method on their own data. Additional datasets provided by the community will be incorporated in a future revision of this manuscript.

2. MOSAIC

MOSAIC is a language model specialized for prompt-based radiological report classification, available in either 4B and 12B versions. The final models are based on supervised fine tuning of the `MedGemma-4B` and `Gemma-12B` models using publicly available datasets. We describe our design decisions, including model selection, prompt design, fine-tuning strategies, and data augmentation to quantify the impact of each design choice.

2.1. Data

Few public radiological reports datasets are currently available, due to the risk of de-anonymization of patients or clinicians. Most reports are machine-annotated, which leads to noisy labels unsuitable for training language models. For high-quality results, we consider only datasets manually annotated or checked by radiologists. The key statistics of the datasets are presented in Table 1.

MIMIC-CXR (Johnson et al., 2019) is a collection of chest X-ray English radiological reports. It has annotations for 3 possible types of mentioned finding: positive mention ($+$), when a finding is described as present in the report; negative mention ($-$), as finding is described as absent in the report; and uncertain mention ($\sim$), when a definitive conclusion cannot be reached on a specific finding. All not mentioned findings in the report are assigned a "not mentioned" label. Only a small subset of the dataset was manually annotated and released.

PadChest-GR (Castro et al., 2024) is a manually-curated chest X-ray dataset. The original dataset was a collection of reports from a hospital in Spain. Later, the sentences describing abnormalities were extracted from text through LLM prompting, and translated to English. The extracted sentences were reviewed by a team of radiologists while inspecting the associated X-ray. We limit the label set to findings that occur at least 30 times.

CASIA-CXR (Metmer and Yang, 2024) is a French chest X-ray dataset. Each report was assigned one of five findings as positively mentioned. Labels for both *PadChest-GR* and *CASIA-CXR* can only be positively mentioned findings.

REFLACX (Bigolin Lanfredi et al., 2022) is based on *MIMIC-CXR*, from which unlabeled reports were extracted for manual annotation in two phases, with two different taxonomies: we refer to these as *REFLACX*^{I} and *REFLACX*^{II}. Each finding in these datasets was annotated as not mentioned, or

Dataset		Language	Modality	Number of Findings	Avg. Chars	Mention Classes	Train	Dev	Test
MIMIC-CXR	M	en	Chest X-Ray	14	760	$+, -, \sim$	535	50	100
PadChest-GR	P	es, en	Chest X-Ray	49	115	$+$	1951	100	879
CASIA-CXR	C	fr	Chest X-Ray	5	400	$+$	7677	100	3334
DanskCXR	D	da	Chest X-Ray	48	312	$+, -$	1600	125	750
Reflacx^{I}	R^{I}	en	Chest X-Ray	14	216	$+, \sim$	68	50	120
Reflacx^{II}	R^{II}	en	Chest X-Ray	15	201	$+, \sim$	1046	52	1098
DanskMRI	B	da	Brain MRI	3	1941	$+, -, \sim$	194	50	345

Table 1: Overview of the datasets used, including language, modality, number of findings, average characters per report, data split, and mention classes, namely positive $(+)$, negative $(-)$, and uncertain $(\sim)$ mentions of findings.

as mention with varying degrees of certainty, with a score from 1 to 5, following the definition by (Panicek and Hricak, 2016). As granular uncertainty detection is not the scope of this study, we map probable findings as positive mentions (score of 4 and 5), and otherwise uncertain.

DanskCXR (Anonymous, 2025) is a private chest X-ray dataset collected from hospitals and clinics in Denmark. Along with *MIMIC-CXR*, it is the only dataset that has annotations for negative mentions of findings. We limit our experiments to the 14 most frequent findings to exclude rare conditions.

We define as *DanskMRI* a dataset of MRI reports from Danish hospitals collected by Beliveau et al. (2024), who focused on identifying epilepsy-related findings in brain MRI radiology reports. MRI is a different medical imaging modality: the style and structure of MRI reports differ significantly from those of chest X-rays, presenting a distinct out-of-distribution scenario.

The dataset splits were generated using stratification (Sechidis et al., 2011) and implemented using the `iterative-stratification` Python library, trying to have a consistent number of validation samples across datasets.

2.2. Model selection and fine-tuning

We select a small language model to base MOSAIC on through an initial set of experiments comparing two model families at different sizes: Llama 3 (3B and 8B) (Grattafiori et al., 2024) and Gemma 3 (4B and 12B) (Team et al., 2025) model families. These models have similar computational requirements and inherent multilingual capabilities. Additionally, we include `Mmed-Llama-8B`, a multilingual medical foundation model based on `Llama-8B` (Qiu et al., 2024). Lastly, we test also three larger models of the same families (`Gemma-27B`, `MedGemma-27B` and `Llama-70B`). While these models are not feasible for local deployment on consumer-grade GPUs due to their high resource requirements, they serve to contextualize the performance of smaller alternatives. We select the instruction-tuned version of each model, given their general superiority for structured outputs (Zhang et al., 2025). Closed-source models are excluded, as they do not meet the crucial privacy requirements of this task.

Each model is fine-tuned on an NVIDIA RTX 4090 GPU (24GB) using its 4-bit quantized form along with Rank-Stabilized LoRA adapters, except for `Gemma-27B` and `Llama-70B` which are trained on a H100 (94GB). `Mmed-Llama-8B` does not have a 4-bit quantized model version. All the prompts and findings sets are written in English to leverage the models' stronger alignment to English instructions. The prompt structure is adapted based on the dataset's label taxonomy to ensure consistency with its definition of positive, uncertain, negative, and unmentioned findings. As an example, the prompt for *CASIA* is:

```
You are a helpful radiology assistant.
Given a radiology report, classify each
abnormality into a class. Output a valid
JSON with each abnormality as key, and the
class as value. The keys must be
['cardiomegaly', 'mass', 'pleural
effusion', 'pneumonia', 'pneumothorax'].
The values can be one of [-1, 1]. The
values have the following interpretation:
(1) the abnormality was mentioned, even
with uncertainty, in the report e.g. 'A
large pleural effusion', 'The cardiac
contours are stable.', 'The cardiac size
cannot be evaluated.'; (-1) the abnormality
was not mentioned in the report, or the
abnormality was negatively mentioned in the
report; e.g. 'No pneumothorax.'.
```

Fine-tuning is conducted using the Unsloth library (Han et al., 2023), while inference is performed with vLLM (Kwon et al., 2023). Before inference, the LoRA adapters are merged into the base models in 16-bit precision. Cross-entropy loss is used as the objective function. Fine-tuning and sampling parameters are selected based on empirical testing and a random hyperparameter search conducted using Weights & Biases (Biewald, 2020). Train-

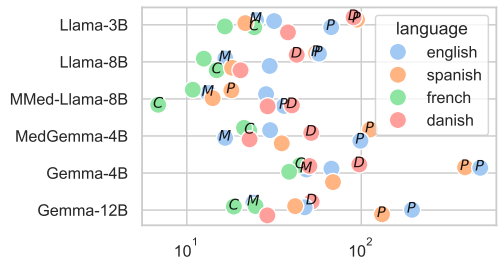


Figure 1: Perplexity scores on the SIB-200 and the chest X-ray reports, with the latter denoted by the dataset's initial.

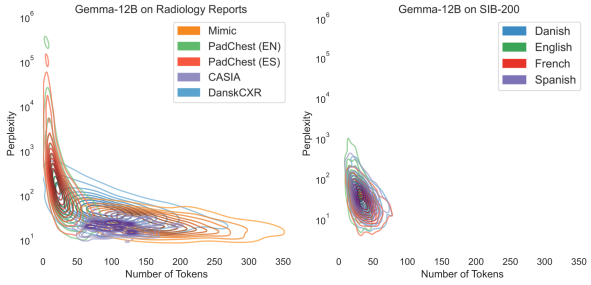


Figure 2: Distribution of perplexity scores as a function of sentence length on radiological reports and SIB-200, on model `Gemma-12B` .

ing configurations and hyperparameters are documented in the accompanying code repository.

2.3. Metrics

We use the F1 score to evaluate the extraction of findings for positive and negative mentions. The average of these scores yields the macro F1 score, reported as $(+)F1$ and $(-)F1$, respectively. To mitigate the impact of class imbalance, we report the weighted F1 score as $(w)F1$, which incorporates class support into the macro F1 calculation.

3. Results

To evaluate a model's ability to classify the radiology reports across multiple languages and taxonomies, we look at two key aspects: language competency and task competency.

3.1. Language competency

Language competency is the ability of a model to understand and generate text in different languages. It reflects how the model captures linguistic patterns and vocabulary, often measured by metrics like perplexity. Strong language competency enables better performance across multilingual tasks. SIB-200 is a large-scale, open-source benchmark for topic classification in 200 languages (Adelani et al., 2024). To assess language modeling capabilities, Figure 1 shows the distribution of perplexity scores on SIB-200 across the four focus languages in this paper and the chest X-ray reports. *MIMIC* and *CASIA*, along with their language counterparts in SIB-200, achieve lower perplexity scores than *DanskCXR* and *PadChest*. While Danish and Spanish datasets in SIB-200 generally yield higher perplexity, the radiology datasets exhibit particularly high scores. $PadChest_{en}$ also records higher perplexity than other English datasets. Among the models, `Mmed-Llama-8B` consistently attains the lowest perplexity, whereas the Gemma family yields the highest values, with `Gemma-4B` performing notably poorly on *PadChest*and `MedGemma-4B` showing a smaller spread. Figure 2 illustrates the relationship between text length and perplexity for `Gemma-12B`. SIB-200 shows a similar distribution, which does not seem affected by language. Source corpus strongly influences the perplexity score: in particular on radiological reports, *MIMIC* and $PadChest_{en}$, have very different distributions, even if both are written in English, confirming how language has a lower impact than dataset provenance in this task. Similar results were observed for other models but are omitted for brevity.

> ***Finding* 1.** Larger models yield lower perplexity, with domain-adapted variants providing additional gains. Dataset provenance also matters, with models performing better on data aligned with their training domain.

3.2. Task competency

Table 2 reports the performance of each model across five chest X-ray datasets under three experimental settings, evaluated using the $(+)F1$ score: zero-shot (ZS), where the model is directly prompted to classify the text according to the given taxonomy; three-shot (3S), using three examples randomly drawn from the corresponding training sets; and fine-tuned (FT) on *MIMIC*. In the 3-shot setting, three examples from the respective training set are provided. When fine-tuning, the models were instructed to output a valid JSON with each abnormality as key, and the mention type as the value, using a zero-shot prompt. The outputs are then parsed and validated.

A clear trend emerges across all models and datasets: performance improves with 3-shot prompting. Even if fine-tuning is performed only on one dataset, some models achieve an improvement over the other tested datasets. `Gemma-12B` consistently leads in performance, often achieving the highest F1 scores across multiple datasets and settings, particularly excelling in 3S and FT scenarios. `MedGemma-4B` also demonstrates strong

Dataset	MIMIC			PadChest$_{en}$			PadChest$_{es}$			CASIA			DanskCXR			*Average*		
Experiment	ZS	3S	FT	ZS	3S	FT	ZS	3S	FT	ZS	3S	FT	ZS	3S	FT	ZS	3S	FT
`Llama-3B`	46	53	77	60	66	67	39	53	53	55	75	79	53	53	57	50	60	66
`Llama-8B`	54	61	86	77	76	78	69	67	71	70	75	77	61	63	65	66	68	75
`Mmed-Llama-8B`	39	52	82	42	75	78	29	62	68	60	67	77	45	59	54	43	63	72
`Gemma-4B`	60	62	86	68	75	74	62	72	67	69	73	80	43	64	60	60	69	73
`MedGemma-4B`	55	59	88	61	77	74	53	72	71	69	82	82	62	65	58	60	69	75
`Gemma-12B`	65	70	84	79	79	80	76	76	76	76	76	81	69	75	69	73	75	78
`Gemma-27B` $*$	68	69	87	81	81	82	81	82	82	75	77	81	71	75	68	75	77	80
`MedGemma-27B` $*$	70	69	87	81	83	83	81	84	82	76	78	82	72	76	70	76	78	81
`Llama-70B` $*$	69	72	81	78	79	78	74	70	78	68	79	79	68	73	67	71	75	77

Table 2: Classification performance of language models on chest X-ray radiological free-text reports, measured in $(+)F1$ and ordered by model family and size. Models are tested under three settings: zero-shot (ZS); three-shot (3S) with three examples drawn from corresponding training sets; and the model fine-tuned on *MIMIC* (FT). We indicate with $(*)$ models fine-tuned on a 94GB H100, instead of a consumer-grade 24GB RTX 4090.

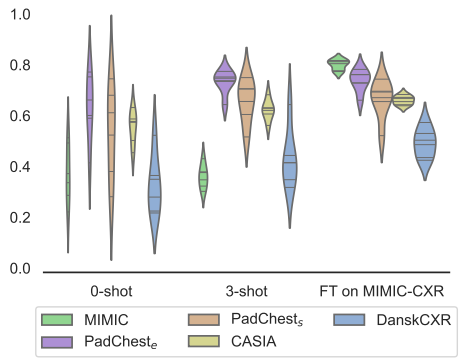


Figure 3: Distribution of $(+)F1$ scores from Table 2. Performance improves with increased supervision, but transferability varies by dataset.

performance, especially considering its smaller size, showing competitive F1 scores and even surpassing larger models in specific instances. `Llama-3B` and `Mmed-Llama-8B` generally show lower performance compared to Gemma variants, while `Llama-8B` performs generally well. Surprisingly, `Mmed-Llama-8B` performs poorly, especially compared to `Llama-8B`, particularly in zero-shot.

On the *MIMIC* dataset, CheXbert reports radiologist-level performance with a $(w)F1$ of 0.809, compared to 0.743 for rule-based methods and 0.798 for BERT-based classifiers (Smit et al., 2020). On a comparable subset, `MedGemma-4B` achieves $(w)F1$=88, suggesting that our approach can achieve expert-level performance. Large models (27B and 70B) have much higher scores on ZS and 3S, compared to the smaller ones. When fine-tuned, the performance is similar or equal on the target set, but these models better generalize on the test sets unseen during fine-tuning.

Figure 3 illustrates the distribution of the scores. While 0-shot learning shows a wide variability, the distributions become significantly tighter and higher for 3-shot learning. This trend is further pronounced when models are fine-tuned on *MIMIC*, suggesting a more robust and reliable performance. Interestingly, models consistently perform better on *PadChest$_{en}$* than on *PadChest$_{es}$*, despite both datasets containing the same reports but in different languages. This suggests that the task is easier in English than in Spanish. We conclude that larger models offer a good solution when no annotated data is available, but lack the accuracy of fine-tuning on the target data. Remarkably, in the fine-tuned setting smaller models match their performance, but show reduced generalization to out-of-distribution datasets. Table 3 shows runtime for training experiments on MIMIC, that we can compare to previous classification performance. Between the smaller models, `Gemma-12B` exhibits superior F1 scores but has higher memory and runtime demands. Additionally, training or inference may fail due to limited resources with large prompts (e.g. when the findings set or report is too long). `MedGemma-4B` stands out for its efficiency and competitive performance.

> ***Finding* 2.** In-context examples and supervised fine-tuning consistently improves performance compared to direct prompting. The best performing small-scale models are `MedGemma-4B` and `Gemma-12B`.

3.3. Multi-dataset fine-tuning

Table 4 reports $(+)F1$ scores using a series of increasingly comprehensive training setups, on the best-performing models in Section 3.2. Each successive row in the table reflects an incremental expansion of the training data, each indicated by the

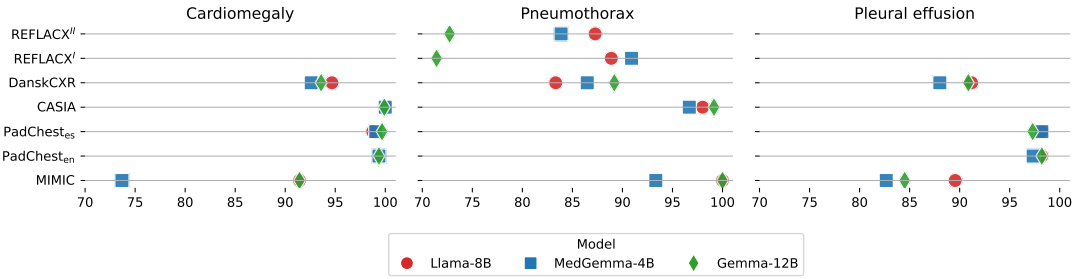


Figure 4: Performance as $(+)F1$ score of `MedGemma-4B`, `Llama-8B`, and `Gemma-12B` fine-tuned on MP$_{E+S}$C, with detailed results on three key Chest X-ray pathologies: Cardiomegaly, Pneumothorax, and Pleural Effusion. The data is presented across a range of chest X-ray datasets, illustrating model-specific and dataset-specific performance.

	Memory (GB)	Runtime (m)
`Llama-3B`	4.70	9.66
`Mosaic-4B`	8.69	13.98
`Gemma-4B`	9.16	13.88
`Llama-8B`	9.61	20.68
`Gemma-12B`	16.24	33.64
`Mmed-Llama-8B`	19.01	19.33
`Gemma-27B`	43.85	36.58
`MedGemma-27B`	41.48	34.65
`Llama-70B`	78.14	79.27

Table 3: Training runtime (in minutes) and peak GPU memory usage for models trained on *MIMIC*.

initial of the included dataset. The MP$_E$ setting includes only English-language data from *MIMIC* and *PadChest*$_{en}$, serving as the monolingual baseline. MP$_{E+S}$ adds Spanish samples, introducing cross-lingual variation. In MP$_{E+S}$C, French data from CASIA is incorporated, followed by the final setting MP$_{E+S}$CD, which adds Danish samples from DanskCXR, creating a multilingual, multi-institutional training configuration. We observe that the model trained under the MP$_E$ condition, using only English data, still achieves strong performance on *PadChest*$_{es}$, which contains the same reports as P$_E$ in Spanish. This result suggests that the model can generalize to the task itself without direct exposure to Spanish during fine-tuning, providing evidence of emerging *task competency*, even when the training data is exclusively monolingual, as long as the clinical structure remains consistent.

While the inclusion of additional languages and datasets does lead to gradual performance improvements, these gains are most pronounced in the final configuration MP$_{E+S}$CD, where all sources are present. The inclusion of *CASIA* in MP$_{E+S}$C yields near-to-perfect results on *CASIA*. This is not surprising, as this dataset is the largest (7.6k samples, compared to 0.5k of *MIMIC*) and with the simplest task, multi-class classification over 5 findings. Including *CASIA* and *DanskCXR* slightly degrades performance on *MIMIC* in smaller models, likely due to *MIMIC* becoming less proportionally represented in the combined training data. This effect is not observed in `Gemma-12B`. Figure 4 further illustrates this trend for three critical pathologies, observed across the relevant training sets and the English-language datasets *REFLACX*^{I} and *REFLACX*^{II}, under the MP$_{E+S}$CD setting. While `Gemma-12B` achieves the highest F1 scores, `MedGemma-4B` performs on par with or better than larger models in several scenarios, and on out-of-distribution *REFLACX* datasets.

> ***Finding* 3.** Adding more datasets improves performance, particularly in multilingual training. English-only models transfer to Spanish via shared clinical structure.

3.4. Taxonomy adaptation

Table 5 looks at the performance of `MOSAIC` trained on MP$_{E+S}$CD to assess task competency. The left columns (M and P$_E$) reflect the model's initial task competency on English-language findings. The red "X" marks in these columns indicate that these specific findings are *not present* in that dataset taxonomy. The right columns assess the model's task competency by measuring its ability to generalize to unseen English datasets R^{I}and R^{II}, both before and after fine-tuning on new data ($\Rightarrow$R). Although "Consolidation" is present in all datasets, fine-tuning on new data still yields substantial improvements. On findings that are present in only one of the training sets ("Nodule" and "Pneumothorax"), fine-tuning also significantly improves the models' ability to adapt their existing task competency to new distributions. However, for findings *not* included in the initial training set taxonomy, the fine-tuned model shows a small improvement ("Enlarger Hilum", of which there are 29 samples in the whole dataset) or, in the case of "Emphysema" (N=6), fine-tuning hurts performance.

	MIMIC			PadChest$_{en}$			PadChest$_{es}$			CASIA			DanskCXR			*Average*		
Datasets	MedGemma-4B	L1ama-8B	Gemma-12B	MedGemma-4B	L1ama-8B	Gemma-12B	MedGemma-4B	L1ama-8B	Gemma-12B	MedGemma-4B	L1ama-8B	Gemma-12B	MedGemma-4B	L1ama-8B	Gemma-12B	MedGemma-4B	L1ama-8B	Gemma-12B
M	87	86	84	74	78	80	71	71	76	82	77	81	58	65	69	74	75	78
MP$_E$	78	**89**	85	92	**95**	**95**	84	85	91	83	74	80	65	65	70	80	82	84
MP$_{E+S}$	80	82	84	93	94	**95**	92	**94**	**94**	83	76	79	61	61	69	82	81	84
MP$_{E+S}$C ⋆	76	85	84	92	**95**	94	89	**94**	**94**	**99**	**99**	**99**	66	63	69	84	87	88
MP$_{E+S}$CD	76	76	85	93	93	**95**	91	92	**94**	**99**	**99**	**99**	82	84	**86**	88	89	92

Table 4: $(+)F1$ scores with incrementally expanded multilingual training configurations on chest X-ray radiological reports. In bold, the best result over each dataset. MP$_E$ establishes a monolingual English baseline. Subsequent rows incrementally incorporate radiological reports in Spanish (P$_S$), French (C), and Danish (D), culminating in a multilingual, multi-institutional training setup. `MedGemma-4B` is 3 times smaller than `Gemma-12B`, but it still achieves a competitive performance. The symbol ⋆ indicates the training configuration used for the open-source `MOSAIC-4B` and `MOSAIC-12B`.

M	P$_E$		R$^{\text{I}}$	⇒R$^{\text{I}}$	R$^{\text{II}}$	⇒R$^{\text{II}}$
77	100	Consolidation	67	91	65	92
93	✗	Pneumothorax	90	100	84	93
✗	68	Nodule	47	60	✗	✗
✗	100	Hiatal Hernia	✗	✗	89	100
✗	✗	Emphysema	71	50	✗	✗
✗	✗	Enlarged Hilum	✗	✗	74	75

Table 5: Taxonomy adaptation in English of `MOSAIC` trained on MP$_{E+S}$CD, measured in $(+)F1$. Left columns show performance on present findings in training sets M and P$_E$; right columns show generalization to unseen datasets R$^{\text{I}}$ and R$^{\text{II}}$ before and after fine-tuning (⇒R).

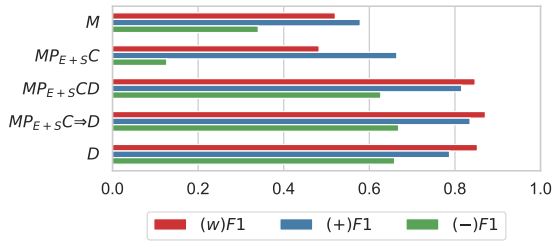


Figure 5: Evaluation of different pretraining and fine-tuning configurations of `MOSAIC` on the *DanskCXR*. Fine-tuning directly on D dataset outperforms zero-shot and task pretraining.

***Finding* 4.** Fine-tuning improves generalization to unseen datasets, especially for findings present in the training data, but offers limited gains for rare or missing findings.

3.5. Domain adaptation

DanskCXR is challenging out-of-distribution (OOD) training set for evaluating model adaptation beyond in-domain performance. This dataset is complex, as it is multi-class, multi-label, sourced from multiple institutions, and written in Danish. Notably, the current SOTA on this task is a BERT-based model that achieves a $(+)F1$ score of 88, across both positively and negatively mentioned findings. Figure 5 shows how different training strategies impact performance on *DanskCXR*. When pretrained on M or on MP$_{E+S}$C, the model struggles with negative finding detection: $(-)F1$=34, $(-)F1$=13, respectively. Incorporating *DanskCXR* into the pretraining phase yields no notable improvement over simply fine-tuning on *DanskCXR* alone. However, adapting a fine-tuned `MOSAIC` to *DanskCXR* (MP$_{E+S}$C⇒D) provides a modest performance boost, achieving $(+)F1$=84. We next ask: what is the minimal data requirement to match full-dataset performance on *DanskCXR*? As shown in Figure 6, we find that using as little as 30% of the dataset (480 examples) is sufficient to match the performance achieved with the full training set. To further examine data efficiency of small LLMs, Table 6 presents results when training is restricted to just 5% of the dataset (80 annotated examples).

***Finding* 5.** Domain adaptation performance is maximized by sequential adaptation to the target domain dataset. Direct pretraining on mixed datasets offers limited gains.

***Finding* 6.** Domain adaptation reaches its maximum performance with only 30% of the expert-annotated data (480/1600 examples).

	5%	10%	100%
D	75	80	85
D_e	73	74	85
D_{e+d}	79	80	86
$MP_{E+S}C \Rightarrow D$	75	82	87
$MP_{E+S}C \Rightarrow D_{e+d}$	82	84	86

Table 6: $(w)F1$ scores of MOSAIC trained on *DanskCXR* (D) and its English translation (D_e), with and without pretraining on $MP_{E+S}C$, on full data (100%) and two training dataset subsets (5% and 10%).

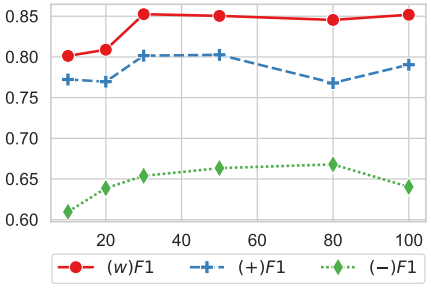


Figure 6: Data ablation study on *DanskCXR* shows that using just 480 examples (30% of the full dataset) is sufficient to reach peak performance.

3.6. Data augmentation

We also explore the benefit of data augmentation in domain adaptation. This experiment involves machine translating the original Danish dataset into English, obtained using `Gemma-27B`, to which we refer to as D_e. Models trained on this data were evaluated on the translated test set, which consistently yielded stronger results. Remarkably, initializing MOSAIC from $MP_{E+S}C$ and fine-tuning on only 5% of D achieves a $(w)F1$ score just 4 points below that of the full-data setup, highlighting the impact of pretraining and the surprisingly low data requirement for effective adaptation with data augmentation.

***Finding* 7.** Data augmentation through machine translation improves classification performance when there are limited resources in a desired language. This can reduce the effort required for expert data annotation.

3.7. Imaging modality adaptation

The *DanskMRI* dataset consists of Danish MRI reports annotated for three epilepsy-related brain abnormalities. Unlike chest X-ray datasets, these findings relate to neurological imaging, introducing both clinical and linguistic shifts. As shown in Figure 7, adaptive fine-tuning on external chest X-ray datasets ($MP_{E+S}C$) improves performance

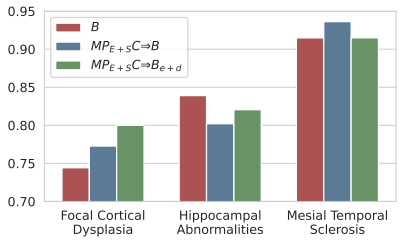


Figure 7: Performance of MOSAIC on the *DanskMRI* dataset, measured as $(+)F1$ across three epilepsy-related abnormalities from MRI reports.

over the base model for Focal Cortical Dysplasia and Mesial Temporal Sclerosis. This suggests that while cross-modality transfer can be effective, it may not generalize uniformly across all conditions. Adding English data augmentation improves consistency across all findings. In particular, it recovers performance on Hippocampal Abnormalities without sacrificing gains on the others. These results highlight the benefit of lightweight augmentation when adapting to new modalities, especially under language and data constraints, as only 194 examples are provided for fine-tuning.

***Finding* 8.** Cross-imaging modality transfer improves performance; English data augmentation enhances consistency.

4. Conclusion

We present MOSAIC, an approach for radiology report classification designed for local deployment and practical use. MOSAIC is a fine-tuned small language model trained on publicly available multilingual chest X-ray datasets (*MIMIC*, *PadChest*, *CASIA*). Unlike existing methods, it remains efficient while being flexible across languages and label taxonomies. We evaluate MOSAIC on English, Spanish, French, and Danish reports across two imaging modalities, finding robust performance in all settings. With as few as 80 annotated examples, it adapts effectively to new data distributions. As the first method to combine efficient multilingual adaptation with taxonomy-agnostic prompting, MOSAIC delivers state-of-the-art results across languages and taxonomies on standard hardware. By removing the need for large, expert-annotated datasets and language-specific models, it offers an accessible, scalable alternative for clinical NLP. MOSAIC is released in 4B and 12B parameter versions, both trained on public data, to support responsible and reproducible AI in healthcare.

Limitations

While MOSAIC shows strong performance across a range of datasets, it has several limitations. First, the model lacks interpretability: its predictions are based on internal transformer representations without transparent reasoning or token-level explanations. This opacity can be a barrier in clinical contexts, where trust and accountability are critical. Future work could explore integrating explainable components. Second, for some datasets there are no established baselines, making it challenging to benchmark relative performance. Third, our evaluation focuses primarily on positively mentioned findings. However, some clinical applications may require precise handling of negative and uncertain statements, which we do not evaluate on. Finally, although MOSAIC generalizes well across languages and imaging modalities, it struggles with rare or unseen conditions.

Acknowledgments

This research was funded by Innovation Fund Denmark grant number 0176-00013B and by Lundbeck Foundation grant (R279–2018–1145). Ethical approval for the dataset *DanskCXR* was obtained on 11 May 2022 from the Regional Council for Region Hovedstaden (R-22017450). Ethics approval for data collection for *DanskMRI* has been granted by the relevant ethics committees (CVK-2112460) and approval for data processing has been given by the Danish Data Protection Authorities.

5. Bibliographical References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2024. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects.

Fares Al Mohamad, Leonhard Donle, Felix Dorfner, Laura Romanescu, Kristin Drechsler, Mike P. Wattjes, Jawed Nawabi, Marcus R. Makowski, Hartmut Häntze, Lisa Adams, Lina Xu, Felix Busch, Aymen Meddeb, and Keno Kyrill Bressem. 2025. Open-source large language models can generate labels from radiology reports for training convolutional neural networks. *Academic Radiology*, 32(5):2402–2410.

Anonymous. 2025. Effective machine learning techniques for non-english radiology report classification: A danish case study. *AI*, 6(2).

Vincent Beliveau, Helene Kaas, Martin Prener, Claes N Ladefoged, Desmond Elliott, Gitte M Knudsen, Lars H Pinborg, and Melanie Ganz. 2024. Classification of radiological text in small and imbalanced datasets in a non-english language. *arXiv preprint arXiv:2409.20147*.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F Auffermann, Jessica Chan, Phuong-Anh T Duong, Vivek Srikumar, Trafton Drew, Joyce D Schroeder, and Tolga Tasdizen. 2022. Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific data*, 9(1):350.

Daniel C Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, et al. 2024. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *arXiv preprint arXiv:2411.05085*.

Jaime Collado-Montañez, María-Teresa Martín-Valdivia, and Eugenio Martínez-Cámara. 2025. Data augmentation based on large language models for radiological report classification. *Knowledge-Based Systems*, 308:112745.

Felix J Dorfner, Liv Jürgensen, Leonhard Donle, Fares Al Mohamad, Tobias R Bodenmann, Mason C Cleveland, Felix Busch, Lisa C Adams, James Sato, Thomas Schultz, et al. 2024. Is open-source there yet? a comparative study on commercial and open-source llms in their ability to label chest x-ray reports. *arXiv preprint arXiv:2402.12298*.

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*.

Jawook Gu, Han-Cheol Cho, Jiho Kim, Kihyun You, Eun Kyoung Hong, and Byungseok Roh. 2024. Chex-gpt: Harnessing large language models for enhanced chest x-ray report labeling. *arXiv preprint arXiv:2401.11505*.

Daniel Han, Michael Han, and Unsloth team. 2023. Unsloth. http://github.com/unslothai/unsloth. Software.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik

Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Alistair Johnson, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet*, 101:215–220.

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*.

Hidetoshi Matsuo, Mizuho Nishio, Takaaki Matsunaga, Koji Fujimoto, and Takamichi Murakami. 2024. Exploring multilingual large language models for enhanced tnm classification of radiology report in lung cancer staging. *Cancers*, 16(21):3621.

Markus Mergen, Daniel Spitzl, Conrad Ketzer, Maximilian Strenzke, Alexander W. Marka, Marcus R. Makowski, Keno K. Bressem, Lisa C. Adams, and Florian T. Gassert. 2025. Leveraging large language models for accurate ao fracture classification from ct text reports. *Journal of Imaging Informatics in Medicine*.

Hichem Metmer and Xiaoshan Yang. 2024. An open chest x-ray dataset with benchmarks for automatic radiology report generation in french. *Neurocomputing*, 609:128478.

Luc Mottin, Jean-Philippe Goldman, Christoph Jäggli, Rita Achermann, Julien Gobeill, Julien Knafou, Julien Ehrsam, Alexandre Wicky, Camille L Gérard, Tanja Schwenk, et al. 2023. Multilingual recist classification of radiology reports using supervised learning. *Frontiers in digital health*, 5:1195017.

Thao Nguyen, Tam M Vo, Thang V Nguyen, Hieu H Pham, and Ha Q Nguyen. 2022. Learning to diagnose common thorax diseases on chest radiographs from radiology reports in vietnamese. *Plos one*, 17(10):e0276545.

Matteo Olivato, Luca Putelli, Nicola Arici, Alfonso Emilio Gerevini, Alberto Lavelli, and Ivan Serina. 2024. Language models for hierarchical classification of radiology reports with attention mechanisms, bert, and gpt-4. *IEEE Access*, 12:69710–69727.

David M Panicek and Hedvig Hricak. 2016. How sure are you, doctor? a standardized lexicon to describe the radiologist's level of certainty. *American Journal of Roentgenology*, 207(1):2–3.

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. *Nature Communications*, 15(1):8384.

Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. 2024. A scoping review of large language model based approaches for information extraction from radiology reports. *npj Digital Medicine*, 7(1).

Eduardo P Reis, Joselisa PQ De Paiva, Maria CB Da Silva, Guilherme AS Ribeiro, Victor F Paiva, Lucas Bulgarelli, Henrique MH Lee, Paulo V Santos, Vanessa M Brito, Lucas TW Amaral, et al. 2022. Brax, brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pages 145–158. Springer.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *CoRR*, abs/2004.09167.

Kazufumi Suzuki, Hiroki Yamada, Hiroshi Yamazaki, Goro Honda, and Shuji Sakai. 2024. Preliminary assessment of tnm classification performance for pancreatic cancer in japanese radiology reports using gpt-4. *Japanese Journal of Radiology*, 43(1):51–55.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Alessandro Wollek, Sardi Hyska, Thomas Sedlmeyr, Philip Haitzer, Johannes Rueckel, Bastian O. Sabel, Michael Ingrisch, and Tobias Lasser. 2024. German chexpert chest x-ray radiology report labeler. *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, 196(09):956–965.

Eric Yang, Matthew D Li, Shruti Raghavan, Francis Deng, Min Lang, Marc D Succi, Ambrose J Huang, and Jayashree Kalpathy-Cramer. 2023. Transformer versus traditional natural language processing: how much data is enough for automated radiology report classification? *The British Journal of Radiology*, 96(1149):20220769.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2025. Gpt4roi: Instruction tuning large language model on region-of-interest. In *European Conference on Computer Vision*, pages 52–70. Springer.

Yirong Zhou, Paul K Amundson, Fangsheng Yu, Matthew M Kessler, Tammie L S Benzinger, and Franz J Wippold. 2014. Automated classification of radiology reports to facilitate retrospective study in radiology. *Journal of Digital Imaging*, 27(6):730–736.

MedNormJ: A Benchmark Dataset for Medical Concept Normalization in Japanese Clinical Documents

Yuki Tashiro, Seiji Shimizu, Tomohiro Nishiyama, Shoko Wakamiya, Eiji Aramaki
Nara Institute of Science and Technology, Japan
tashiro.yuki.ty3@naist.ac.jp
{shimizu.seiji.so8, nishiyama.tomohiro.ns5, wakamiya, aramaki}@is.naist.jp

Abstract

Medical concept normalization in clinical text is a fundamental technology for the secondary use of clinical data. However, constructing annotated resources for this task is challenging because annotation is both expertise-intensive and methodologically complex. As a result, a standard evaluation dataset for Japanese has yet to be established. In this study, we introduce a Japanese dataset for medical concept normalization, **MedNormJ** , which will be publicly available. The dataset consists of 397 pairs of medical expressions and their corresponding normalized disease names, manually curated from 96 medical documents, including case reports and radiology reports. Furthermore, we conduct comparative experiments using existing normalization approaches to benchmark their performance on this dataset in terms of both accuracy and computational efficiency. Through these experiments, we clarify the present performance level and identify remaining challenges specific to Japanese medical concept normalization.

Keywords: Medical Concept Normalization, Entity Linking, Clinical Text Processing, Benchmark Dataset, Large Language Model

1. Introduction

In recent years, there has been a growing global demand to accelerate drug development and clinical research through the use of real-world data derived from clinical text (Usuyama et al., 2025). Clinical text is a high-density information source that records patients' conditions in detail. However, such text is largely unstructured and exhibits substantial lexical variation, including synonyms, abbreviations, and typographical errors, which hinders large-scale automated analysis. To address this challenge, medical concept normalization has emerged as a foundational technology in clinical natural language processing.

This task can be formulated as mapping disease mentions within medical text (hereafter, mentions) to their corresponding concepts in a standardized terminology set (hereafter, an ontology). The mapped concepts are referred to as normalized forms. As illustrated in Figure 1, when utilizing the ICD-10-based Standard Disease Code Master, a Japanese disease terminology system derived from the International Classification of Diseases (ICD-10)[1], normalization entails selecting the most clinically appropriate concept from a search space of approximately 20,000 candidates for a given mention. For example, mentions such as "well differentiated adenocarcinoma" or "advanced sigmoid colon cancer" would be mapped to the normalized form "sigmoid colon cancer." This process of linking mentions to unique identifiers within ontologies like SNOMED CT (Chang and Mostafa, 2021) or ICD-10 is often referred to as medical coding.

Although the task of normalization is conceptually clear, constructing normalization datasets in the medical domain is challenging for two main reasons. First, annotation is costly: accurately handling complex disease names requires the involvement of experts with medical knowledge, and securing annotators and managing annotation effort pose stronger constraints than in general-domain settings (Wei et al., 2019; Fries et al., 2021). Second, there is uncertainty in deciding normalization granularity. Because medical concepts form deep hierarchical structures in ontologies (Figure 1), it can be difficult to determine which level of the hierarchy a mention should be linked to (Minh et al., 2025). For instance, for a mention such as "highly advanced sigmoid colon cancer," one must decide whether to preserve site information by selecting "sigmoid colon cancer" or to aggregate it into broader concepts such as "colon cancer" or "tumor," depending on downstream analysis goals and context. In other words, medical concept normalization faces two major challenges: a resource challenge, referring to the difficulty of securing expert annotators and a definition challenge, referring to ambiguity in determining the correct normalization, even among experts. These challenges have prevented the establishment of a standard evaluation dataset for Japanese medical text, making it difficult to evaluate normalization methods in terms of both accuracy and computational efficiency.

Given this situation, the development of a carefully designed benchmark dataset becomes essen-

[1] https://www2.medis.or.jp/stdcd/byomei/index.html

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 324–335
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

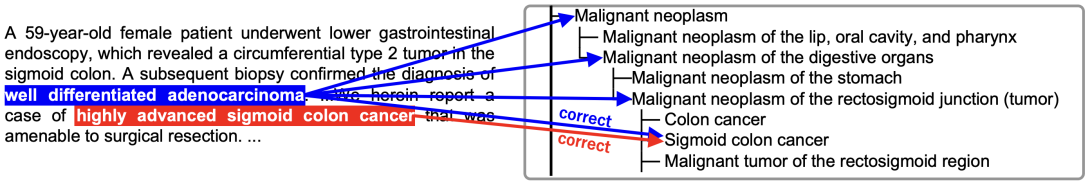


Figure 1: Overview of the medical concept normalization task. In this study, each disease mention in clinical text is mapped to a corresponding ontology concept. Mentions highlighted in red can be normalized literally (e.g., to “sigmoid colon cancer”), whereas those highlighted in blue are ambiguous in isolation and require contextual information for correct normalization. Thus, the task involves context-aware disambiguation based on surrounding text.

Guidelines	INPUT: target term (bold) and its context	OUTPUT: Normalized form and explanation
Principle: Context-dependent specificity When the disease concept to be annotated can be associated with a specific anatomical site (e.g., an organ) or clinical characteristics based on the context, it must be normalized to the **most specific disease name possible**. If the same concept is described at different levels of granularity within a document, the most specific disease name that can be inferred from the context should be selected as the normalized form.	胃底部に隆起性腫瘍を認めた. (An **elevated tumor** was observed in the gastric fundus.)	胃腫瘍 (**Gastric tumor**) Note: As the specific anatomical site is identifiable from the context, the general term “tumor” (腫瘍) must be avoided in favor of the more granular term “gastric tumor”.
Rule 1: Inclusion of lexical modifiers Modifiers that constitute an established medical observation or a standardized clinical concept must be retained in the normalized form.	盲腸に隆起性病変を認めた. (An **elevated lesion** was observed in the cecum.)	隆起性病変 (**Elevated lesion**) Note: The modifier “elevated” is an integral component of the established clinical observation; therefore, it should not be reduced to the more generic “lesion”.
Rule 2: Explicit causative factors When the causative factor is explicitly mentioned, the normalization should include the corresponding clinical category representing that cause.	乳アレルギーがあるため食事制限を行った. (Dietary restriction was implemented due to a **milk allergy**.)	食物アレルギー (**Food allergy**) Note: The specific cause “milk” indicates the category as food-related; therefore, the broader term “allergy” should be avoided.
Rule 3: Pathological diagnosis When a definitive clinical diagnosis is described in pathological findings, the diagnosis should be adopted as the normalized form.	病理所見は低分化型腺癌であった… 上行結腸癌の1例を経験した. (Pathology revealed a **poorly differentiated adenocarcinoma** … a case of ascending colon cancer.)	上行結腸癌 (**Ascending colon cancer**) Note: The final clinical diagnosis takes precedence over the histological description “poorly differentiated adenocarcinoma”.
Rule 4: Abbreviation expansion All medical abbreviations must be expanded to their full, stardardized technical expressions.	既往に**GERD**がある. (The patient has a history of **GERD**.)	胃食道逆流症 (**Gastroesophageal reflux disease**) Note: Abbreviations must be expanded to their full standardized form.
Exception 1: Abstruction of multiple findings When multiple specific findings are listed and a higher-level concept encompasses them, the higher-level concept should be adopted.	胸部X線で小結節，すりガラス影，腫瘤影を認めた. (Chest X-ray revealed **small nodules, ground-glass opacities, and mass shadows**.)	胸部異常陰影 (**Abnormal chest shadow**) Note: Multiple findings are grouped under a single encompassing concept; therefore, individual findings like “small nodule” are not selected.
Exception 2: Unification of synonymous findings Synonymous or closely related terms referring to the same clinical entity (e.g., mass-like pulmonary lesions) should be normalized to a single representative term.	肺に腫瘍／腫瘤性病変を認めた. (A **tumor / mass-like lesion** was observed in the lung.)	肺腫瘤 (**Pulmonary mass**) Note: Synonymous expressions should be unified into a representative term; the general term “tumor” is avoided here for consistency.
Exception 3: Malignancy preference If the context suggests malignancy, the normalized form should prioritize “cancer” over neutral terms like “tumor” to reflect the diagnostic intent.	画像上腫瘤性病変を認め，悪性が強く疑われた. (Imaging revealed a **mass-like lesion**, strongly suggestive of malignancy.)	癌 (**Cancer**) Note: Given the strong clinical suspicion of malignancy, the specific term “cancer” is preferred over the neutral “tumor”.

Table 1: Abstract guidelines for medical concept normalization and corresponding annotation examples.

tial. In this study, we use the concept normalization to collectively refer to the process of linking mentions to standardized concepts. We construct a Japanese medical concept normalization evaluation dataset, **MedNormJ**. We first performed pilot annotations by experts and analyzed cases of disagreement between annotators. Based on the analysis, we created annotation guidelines with explicit rules to unify normalization granularity. Specifically, the principle is to refine to more specific concepts whenever site, characteristics, or stage can be identified from context (Table 1). We then annotated the full dataset according to the guidelines and evaluated existing normalization methods in terms of performance and computational efficiency. The constructed dataset is publicly released for research purposes. [2]

[2] https://github.com/sociocom/mednormj

In this study, we make three contributions: (1) constructing **MedNormJ**, a Japanese benchmark dataset for medical concept normalization; (2) designing annotation guidelines to address normalization ambiguity; and (3) benchmarking existing normalization methods with respect to accuracy and computational efficiency.

2. Related Work

This section reviews prior research on medical concept normalization from both the dataset and methodological perspectives.

2.1. Medical Concept Normalization Datasets

A variety of benchmark datasets have been developed for medical concept normalization in English. The NCBI Disease Corpus (Doğan et al., 2014) is one of the most widely used datasets for disease normalization, while BC5CDR (Li et al., 2016) and MedMentions (Mohan and Li, 2019) extend coverage to chemical entities and a broad range of UMLS concepts, respectively. These resources have supported the development of representative normalization approaches, including DNorm (Leaman et al., 2013) and SapBERT (Liu et al., 2021a).

In contrast, the Japanese medical domain still lacks a widely shared benchmark dataset for medical concept normalization. Although several studies have constructed proprietary datasets (Morita et al., 2013; Aramaki et al., 2014), these resources either do not provide normalization annotation guideline or restrict normalization to ICD-10 codes. Because ICD-10 is primarily intended for statistical and administrative use, it may not capture the diverse disease expressions observed in clinical text at clinically sufficient granularity.

Annotation is also an essential component of dataset construction. A critical aspect of the annotation process is the assessment of both the effectiveness of the annotation scheme itself and the extent to which annotators correctly understand and follow the guidelines (Fernandes et al., 2025). The success of annotation largely depends on the clarity, consistency, and comprehensiveness of the documentation, as well as on the annotators' training and familiarity with the scheme (Artstein and Poesio, 2008). Well-developed guidelines, which include explicit definitions and illustrative examples, are indispensable for achieving reliable and accurate annotations (Stubbs and Pustejovsky, 2012).

The validation of an annotation scheme typically involves a combination of pilot annotations, iterative refinement of the guidelines, and qualitative analysis of problematic cases (Fernandes et al., 2025). In the annotation process, it is common practice to collect judgments from multiple annotators for each data instance, a methodology that has been widely recognized as effective for improving annotation quality (Snow et al., 2008). To quantitatively assess annotation quality, inter-annotator agreement (IAA) is the most commonly used metric, as it measures the consistency of annotations across annotators (Artstein and Poesio, 2008). Representative measures for computing IAA include Cohen's κ coefficient (Cohen, 1960), Krippendorff's α coefficient (Krippendorff, 2018), and simple agreement rates, commonly referred to as accuracy. While high IAA scores suggest that the guidelines are clear and effective, low agreement may arise from a variety of factors (Bayerl and Paul, 2024), often revealing underlying ambiguities or conceptual difficulties in the annotation task.

2.2. Medical Concept Normalization Methods

Existing approaches to medical concept normalization can be broadly categorized into dictionary-based, machine learning-based, bi-encoder-based, and large language model (LLM)-based methods.

Dictionary-based methods provide a simple baseline by directly matching an input mention against entries in a controlled vocabulary. String-based matching includes exact match and edit distance measures such as Levenshtein distance (Levenshtein et al., 1966). Although these techniques offer fast inference and low implementation cost, they are sensitive to out-of-vocabulary mentions and struggle to resolve context-dependent ambiguity.

To mitigate these limitations, machine learning–based approaches have been proposed. In particular, learning-to-rank methods such as DNorm (Leaman et al., 2013) model the correspondence between mentions and normalized forms using engineered features, and have been shown to outperform purely string-based baselines. For Japanese, extensions of DNorm have also been explored (Ujiie et al., 2020), where language-specific preprocessing yields additional gains.

More recently, bi-encoder approaches based on pretrained language models (e.g., BERT) have become a mainstream paradigm. These methods embed mentions and candidate normalized forms into a shared vector space and perform normalization by nearest-neighbor retrieval under an embedding similarity metric. Models such as SapBERT (Liu et al., 2021a,b) , which leverage synonym supervision during representation learning, achieve strong performance on biomedical normalization tasks.

LLM-based approaches have recently gained increasing attention. LLMs can incorporate broader contextual information and have been used to rerank candidate lists produced by bi-encoder re-

trieval, often improving accuracy (Abdulnazar et al., 2024). Nevertheless, LLM-based reranking typically incurs substantially higher latency and computational cost, which may limit its applicability in operational settings.

3. MedNormJ

In this section, we present the construction process of **MedNormJ**, a Japanese medical concept normalization dataset. Section 3.1 details the source materials used as the basis of **MedNormJ**, while Section 3.2 reports the pilot annotation and analyzes IAA. Section 3.3 describes the development of annotation guidelines based on the pilot analysis, and Section 3.4 presents the main annotation process along with dataset statistics and IAA.

3.1. Materials

As the basis of **MedNormJ**, we adopt MedTxt, an existing Japanese named entity recognition dataset. MedTxt consists of two subsets: MedTxt-CR (Yada et al., 2022), which comprises 50 case reports, and MedTxt-RR (Nakamura et al., 2022), which comprises 46 radiology reports. Their statistics are shown in Table 3. Case reports are academic articles describing rare clinical cases, whereas radiology reports are written by radiologists who interpret medical images such as CT and MRI and summarize imaging findings and associated diseases. In this study, we newly link each disease mention in these datasets to a standardized disease name in the ICD10-based Standard Disease Code Master, which serves as the ontology for normalization.

3.2. Pilot Annotation

To assess the reliability of medical concept normalization and identify sources of annotation ambiguity, we conducted a pilot annotation and computed IAA. For the pilot annotation, we randomly sampled a total of 100 mentions (50 from each dataset) and asked two Japanese healthcare professionals to annotate them independently.

Table 4 shows that the overall accuracy was 0.570, and disagreements were observed in 43 instances. Analysis of the disagreements revealed systematic mismatches mainly due to differences in concept granularity.

Table 2 presents an error analysis of the pilot annotation and shows the top three error categories that were consistently inconsistently annotated among annotators. It reports representative examples of inter-annotator disagreements and their corresponding counts. In total, 24 of the 43 observed errors were attributable to these three categories. For example, for the mention “abnormal lung field shadow,” surface variation occurred where different synonyms such as “abnormal chest shadow” and “abnormal lung shadow” were assigned; additionally, granularity differences were evident depending on whether context supplementation was applied, such as “mass” vs. “lung mass” (Table 2). These results indicate that medical concept normalization is ambiguous and challenging.

3.3. Guideline Development

Based on the pilot analysis, we developed annotation guidelines to ensure consistent normalization quality and to systematize decision criteria. These rules are shown in Table 1. Our primary rule is to refine to lower-level concepts as much as possible when site, characteristics, or stage can be identified from context.

However, the pilot disagreements included patterns that were difficult to resolve with the refinement principle alone. First, radiology reports often describe multiple findings in parallel; normalizing each finding individually can lead to greater variability across annotators. Indeed, in the pilot annotation, the agreement for MedTxt-RR was 0.360, lower than 0.780 for MedTxt-CR (Table 4). Therefore, we defined an exception rule: if multiple findings can be interpreted as the same condition and there exists a reasonable higher-level concept that subsumes them, we normalize to that higher-level concept (Exception 1).

Second, we observed systematic variation arising from synonymous or closely related expressions that refer to the same clinical entity. For example, expressions such as “tumor” and “mass-like lesion” were sometimes normalized differently despite referring to essentially the same concept in context. To reduce variability caused by lexical differences, we introduced a unification rule (Exception 2): synonymous or closely related findings should be normalized to a single representative term.

Third, we observed ambiguity not only in hierarchical granularity but also in diagnostic interpretation. In particular, for expressions where malignancy is not explicitly specified (e.g., “tumor” or “mass”) versus expressions that clearly indicate malignancy (e.g., “cancer” or “adenocarcinoma”), annotators differed in whether to interpret the mention as a neutral lesion concept or as a malignant disease concept. This distinction is not purely hierarchical, but reflects differences in clinical interpretation of malignancy. Therefore, we introduced a term-specific rule (Exception 3): if malignancy is explicitly indicated, normalize to “cancer” or “malignant neoplasm”; otherwise, normalize to “tumor.”

3.4. Main Annotation

We constructed the final evaluation dataset using all available documents, based on the developed

Annotator H1	Annotator H2	Freq.
腫瘤 (Tumor)	肺腫瘤 (Pulmonary mass)	9
肺野異常陰影 (Abnormal lung field shadow)	胸部異常陰影 (Abnormal chest shadow)	8
肺野異常陰影 (Abnormal lung field shadow)	肺部異常陰影 (Abnormal pulmonary shadow)	7

Table 2: Representative examples of systematic disagreements between two annotators (H1 and H2) in the pilot phase. Each row shows normalized forms independently assigned by Annotators H1 and H2 to the same mention in context, along with their frequency. These disagreements mainly reflect differences in normalization granularity and synonym selection.

	CR	RR	MedNormJ
Reports	50	46	96
Mentions	270	127	397
Unique mentions	234	62	295
Unique normalized forms	187	24	206
Mentions/normalized ratio	1.4	5.3	1.9

Table 3: Statistics of MedNormJ (H1&H2 agreement subset). CR and RR correspond to MedTxt-CR and MedTxt-RR, respectively.

Phase	Metric	MedTxt-CR	MedTxt-RR	Overall
Pilot	Accuracy	0.780	0.360	0.570
	Cohen's κ	0.776	0.346	0.565
Main	Accuracy	0.761	0.784	**0.768**
	Cohen's κ	0.759	0.722	**0.762**

Table 4: IAA in the pilot and main phases. Accuracy and Cohen's κ are shown for MedTxt-CR, MedTxt-RR, and overall. Agreement improved in the main phase after guideline refinement.

guidelines. Dataset statistics are shown in Table 3. The dataset covers a total of 96 medical documents, containing 397 normalized mentions in total. The number of unique normalized forms is 206.

We recomputed IAA on the constructed dataset using simple accuracy and Cohen's κ coefficient. As a result, the overall agreement for **MedNormJ** improved from 0.570 to 0.768. Note that the pilot annotation (n=100) and the main annotation (n=515) involve different sample sizes, so the comparison is not strictly controlled. According to commonly used interpretation criteria (Landis and Koch, 1977), this level of agreement corresponds to substantial agreement, indicating that the constructed dataset exhibits a reasonable level of annotation reliability.

On the other hand, systematic disagreements for certain terms were also observed (Table 5). For example, for expressions containing "muscle tone," annotator H1 tended to assign "muscle tone disorder," whereas annotator H2 tended to assign "spasticity." For "atelectasis," disagreements arose regarding whether to normalize to a lower-level concept with site information (e.g., "lower-lobe atelectasis") based on context, or to use the surface form "atelectasis." These results suggest the need for more detailed guidelines for clinically ambiguous expressions, or flexible evaluation schemes such as allowing multiple labels.

4. Experiments

This section summarizes our experimental setup and results on **MedNormJ**. Since the compared methods are described in detail in the subsection below, we focus on the overall evaluation design, including the evaluation metrics and how the results are organized.

4.1. Evaluation Setup

We evaluate medical concept normalization on **MedNormJ** from two perspectives: performance and computational efficiency. Performance is measured by Accuracy@1, and efficiency by average inference time per sample.

Accuracy@1 is defined as the proportion of instances for which the top-ranked normalized form exactly matches the gold label. This metric is widely used in prior work on medical concept normalization and entity linking (Kate, 2021; Luo et al., 2020; Liu et al., 2021a). Accuracy@1 is particularly appropriate for this task because normalization is formulated as selecting a single standardized concept from a predefined candidate set, and exact match evaluation directly reflects whether the model successfully identifies the correct concept.

To evaluate computational efficiency, we measure the average inference time per sample and analyze the trade-off between normalization accuracy and processing cost. All experiments were conducted on a single NVIDIA A100-PCIE-40GB GPU with an AMD EPYC 7702P 64-core CPU. GPT-5.2 (model version: `gpt-5.2-2025-12-11`) was accessed via the OpenAI API under default decoding settings (temperature = 0.0), and the reported inference time includes API latency.

Mentions	Normalized form by H1	Normalized form by H2	Freq.
... Muscle tone (increased)...	Muscle tone disorder	Spasticity	4
Atelectasis	Lower lobe atelectasis	Atelectasis	4

Table 5: Representative examples of systematic disagreement between two annotators (H1 and H2), including contextual information for each mention.

Model	Acc. (H1) (n=511)	Acc. (H2) (n=511)	Acc. (H1 & H2) (n=397)	Inference time (H1 & H2) (sec/sample)
Exact-match	0.399	0.411	0.476	2.10×10^{-5}
Levenshtein	0.406	0.418	0.479	4.13×10^{-4}
DNorm-J	0.451	0.466	0.547	1.00×10^{-3}
Bi-encoder	0.464	0.470	0.549	5.15×10^{-3}
Bi-encoder + GPT-5.2	0.521	0.536	0.602	7.77
Bi-encoder FT	0.534	0.540	0.625	6.00×10^{-3}
Bi-encoder FT + GPT-5.2	**0.572**	**0.560**	**0.653**	8.49
GPT-5.2 (zero)	0.311	0.301	0.358	3.83
GPT-5.2 (zero w/ all candidates)	0.466	0.452	0.565	5.59

Table 6: Comparison of normalization performance (Accuracy@1) and inference time. Acc. (H1) and Acc. (H2) denote accuracy when using annotator H1's and annotator H2's labels as gold, respectively. Acc. (H1 & H2) denotes accuracy when using only instances where both annotators agree as gold. Inference time denotes the average processing time per sample.

4.2. Compared Methods

To assess the difficulty of medical concept normalization on **MedNormJ**, we evaluated a set of representative normalization methods covering diverse modeling paradigms. Our objective is not to propose a new normalization model, but to characterize the current performance landscape of existing approaches for Japanese clinical text, considering both accuracy and inference efficiency.

We selected methods ranging from simple dictionary-based baselines to more advanced Bi-encoder-based and LLM-based approaches. Below, we briefly describe each method used in our experiments.

Dictionary-based As dictionary-based approaches, we employed two methods: an exact-match method and a more flexible edit-distance-based method. In the exact-match method, the input mention is matched against mentions registered in the medical terminology dictionary JMED-DICT mini (Nagai et al., 2025), which contains pairs of mentions and their corresponding normalized forms. If an exactly matching mention is found from this dictionary, the corresponding normalized form is returned as the candidate; otherwise, the input mention itself is returned.

The edit-distance-based method addresses typographical errors and orthographic variations by measuring string similarity. Specifically, we compute the Levenshtein distance (Levenshtein et al., 1966) between the input mention and each mentions in the dictionary. The Levenshtein distance is defined as the minimum number of character insertions, deletions, and substitutions required to transform one string into another. In this method, the dictionary mentions with the minimum distance to the input mention is selected, and its corresponding normalized form is returned as the candidate.

DNorm-J DNorm-J is a Japanese adaptation of DNorm (Leaman et al., 2013) that formulates medical concept normalization as a learning-to-rank problem using TF-IDF-based feature representations. In this study, we evaluated DNorm-J using an existing implementation configured for Japanese clinical text, without additional task-specific tuning.

Bi-encoder The Bi-encoder approach independently encodes mentions and normalized forms into a shared vector space and performs normalization based on embedding similarity. We evaluated two variants: a Bi-encoder without task-specific fine-tuning, and a fine-tuned version trained on the **MedNormJ** training data (Bi-encoder FT).

Bi-encoder + LLM Recent studies have reported improved normalization accuracy by reranking candidate lists generated by a Bi-encoder using LLMs with access to contextual information. Following this line of work, we included a Bi-encoder + LLM configuration in which GPT-5.2 reranks the top candidates. While this approach can improve accuracy, it incurs substantially longer inference time and higher computational cost, which we analyze from an operational perspective.

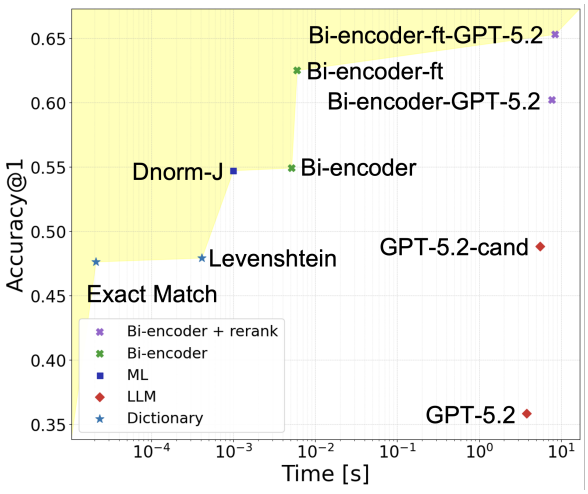


Figure 2: Inference time per sample [s] and Accuracy@1 for each method.

LLM We evaluate two zero-shot prompting strategies for GPT-5.2. The first prompt includes task and output-format instructions, together with the input mention and its context, and requires the model to directly generate the normalized form. The second prompt further provides a candidate list derived from JMED-DICT mini, enabling us to assess the effect of candidate constraints on performance.

4.3. Results

The results are shown in Table 6. While we report performance across all three settings, our analysis primarily focuses on the agreement subset (Acc. (H1 H2)), as it represents the most reliable ground truth derived from annotator agreement. Among the methods we evaluated, the hybrid approach, combining a Bi-encoder fine-tuned on JMED-DICT mini and GPT-5.2-based reranking, achieved the highest Accuracy@1 of 0.653. In contrast, GPT-5.2 in the zero-shot setting yielded the lowest performance, with an Accuracy@1 of 0.358.

Figure 2 plots Accuracy@1 against the average inference time for each method. The horizontal axis shows inference time on a logarithmic scale, and the vertical axis shows Accuracy@1. Methods in the upper-left region of the plot are both faster and more accurate. The yellow frontier line traces, for each latency regime, the maximum Accuracy@1 achieved. Methods on this frontier (exact-match, Levenshtein distance, the fine-tuned Bi-encoder, and the Bi-encoder with GPT-5.2-based reranking) represent the most efficient choices at their respective runtime scales.

In terms of latency, dictionary-based methods were the fastest, requiring only 2.10×10^{-5} seconds per sample. Bi-encoder-based approaches also maintained practical inference times while substantially improving accuracy.

By contrast, LLM-based reranking required over 1 second per sample, indicating much higher latency than dictionary-based methods. Within the reranking framework, using a reasoning-oriented model such as GPT-5.2 also incurred substantially higher computational cost than the other reranking models we evaluated, underscoring the trade-off between accuracy and efficiency.

5. Discussion

5.1. Principal Findings

Future direction: By analyzing expert annotations, we confirmed that medical concept normalization in Japanese medical text inevitably involves variability arising from differences in contextual interpretation and concept granularity. As shown in Table 5, systematic disagreements persist for specific expressions, highlighting the difficulty of defining a single unique correct answer in medical concept normalization. Therefore, future evaluation designs may benefit from more detailed guidelines and flexible frameworks, such as allowing multiple acceptable labels.

Language gap: In the baseline experiments, the highest score was 0.653. In this study, we used SapBERT, which was proposed in previous work (Liu et al., 2021b). The result obtained in our Japanese setting was higher than the previously reported Japanese result for the same model (0.240) on the Japanese subset of the XL-BEL benchmark reported in prior work (Liu et al., 2021b). However, it was still 0.129 lower than the reported English performance (0.782). These results show that performance in Japanese has improved compared to earlier reports, but a clear gap remains between Japanese and English. One main reason for this gap is the difference in the size of the training data. During the pretraining of SapBERT, about one million training instances were available for English, while only about 220,000 were available for Japanese. In addition, our experiments used only about 10,000 Japanese instances. This difference likely limits vocabulary coverage and the variety of expressions in Japanese. Another possible reason is the difference in the evaluation datasets and annotation policies. In Japanese, we observed variation in annotations due to differences in contextual interpretation and concept granularity. As a result, evaluation methods that assume only one correct answer may underestimate the true performance of the model. Overall, these findings suggest that improving performance in Japanese requires not only increasing the amount of training data, for example through synthetic data generation, but also improving the evaluation design and developing training strategies that consider hierarchical relationships

Error Type	Count	Ratio (%)
Prediction of semantically related terms	69	50.0
Hierarchical error (overly abstract prediction)	39	28.3
Mention bias	14	10.1
Hierarchical error (overly specific prediction)	9	6.5
Mentions appearing among candidate normalizations	4	2.9
Others	3	2.1

Table 7: Error types

between medical concepts.

Practical option: Evaluating not only accuracy but also inference time is essential when deploying medical concept normalization systems in real clinical and research settings. Our results with inference time provide practical guidance for selecting appropriate methods in such environments. No single method works best in every situation. The appropriate method depends on the purpose of the system, the required level of accuracy, the acceptable response time, and the available computational resources. When GPUs or other resources needed for deep learning models such as bi-encoders are limited, dictionary-based methods are effective. These approaches are especially suitable for batch processing, where large volumes of data must be handled efficiently. When sufficient computational resources are available and both high accuracy and fast inference per case are required, such as in real-time clinical decision support, a Bi-encoder and LLM are currently the most reliable option.

Our dataset is currently limited to two document types, namely case reports and radiology reports, and thus its coverage is restricted. Because writing styles and the availability of contextual information vary substantially across document types, an important future direction is to expand the dataset to include additional clinical document types, such as progress notes and discharge summaries, thereby connecting normalization research to more practical clinical applications.

Generalizability of exception rules: Although the proposed exception rules improved consistency, their generalizability to larger clinical corpora remains unclear, and additional rules or more flexible frameworks may be required.

5.2. Error Analysis

The most frequent error type involves the prediction of terms that are semantically related to, but not identical with, the gold standard normalized forms, accounting for 69 cases (50.0%). The second most common category consists of hierarchical errors, in which the model predicts a concept that is either more abstract or more specific than the correct one. In particular, overly abstract predictions were observed in 39 cases (28.3%), indicating difficulty in selecting the appropriate level of specificity within the disease name hierarchy.

In addition, 14 errors (10.1%) were attributed to mentions bias, where the model was overly influenced by string-level similarities in the input mention and consequently predicted a contextually inappropriate normalized form with a different clinical meaning. Such errors illustrate the model's tendency to rely on lexical cues at the expense of contextual interpretation.

Overly specific hierarchical errors were observed in 9 cases (6.5%), while errors in which the mention itself appeared among the candidate normalized forms accounted for 4 cases (2.9%). These latter cases suggest that the model failed to sufficiently leverage contextual information, defaulting instead to the mentions even when a more appropriate normalized form existed. The remaining errors (3 cases, 2.1%) were categorized as others.

Overall, this analysis indicates that, despite strong overall performance, the primary sources of error remain challenges in context-dependent decision making and hierarchical concept selection, highlighting directions for future improvements in disease concept normalization models.

6. Conclusion

We constructed a high-quality Japanese medical concept normalization dataset, **MedNormJ**, and reported a detailed annotation process and baseline results. In particular, by developing annotation guidelines that emphasize consistent normalization granularity and contextual interpretation, we achieved high IAA. Our results indicate that medical concept normalization on the dataset presents challenges, particularly due to ambiguity in concept granularity and contextual interpretation, and that there exists a trade-off between performance and computational efficiency across methods. We expect the dataset to contribute to further advances in Japanese medical NLP, including method development, and we plan to expand medical concept normalization datasets in future work.

Acknowledgments

This work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" (Grant Number JPJ012425), JST CREST (Grant Number JPMJCR22N1), and the Japan Society for the Promotion of Science (JSPS) Research Start-up Grant (JP25K24412), Japan.

Bibliographical References

Akhila Abdulnazar, Roland Roller, Stefan Schulz, and Markus Kreuzthaler. 2024. Large language models for clinical text cleansing enhance medical concept normalization. *IEEE Access*.

Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. 2014. Overview of the ntcir-11 mednlp-2 task. In *NTCIR*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2024. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*.

Eunsuk Chang and Javed Mostafa. 2021. The use of snomed ct, 2013-2020: a literature review. *Journal of the American Medical Informatics Association*, 28(9):2017–2026.

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

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Ana Luísa Fernandes, Purificação Silvano, António Leal, Nuno Guimarães, Rita Ribeiro da Silva, Luís Filipe Cunha, and Alípio Jorge. 2025. The incremental process of building an annotation scheme for clinical narratives in portuguese: the contribution of human variation analysis. In *Research Challenges in Information Science-19th International Conference, RCIS 2025, Seville, Spain, May 20-23, 2025, Proceedings, Part II*.

Jason A Fries, Ethan Steinberg, Saelig Khattar, Scott L Fleming, Jose Posada, Alison Callahan, and Nigam H Shah. 2021. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature communications*, 12(1):2017.

Rohit J Kate. 2021. Clinical term normalization using learned edit patterns and subconcept matching: system development and evaluation. *JMIR Medical Informatics*, 9(1):e23104.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

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.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of ACL-IJCNLP 2021*.

Yen-Fu Luo, Sam Henry, Yanshan Wang, Feichen Shen, Ozlem Uzuner, and Anna Rumshisky. 2020. The 2019 n2c2/umass lowell shared task on clinical concept normalization. *Journal of the American Medical Informatics Association*, 27(10):1529–e1.

Dao Sy Duy Minh, Nguyen Lam Phu Quy, Pham Phu Hoa, Tran Chi Nguyen, Huynh Trung Kiet, and Truong Bao Tran. 2025. Dragon: Dual-encoder retrieval with guided ontology reasoning for medical normalization. In *Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association*, pages 230–239.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*.

Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, and Eiji Aramaki. 2013. Overview of the NTCIR-10 MedNLP Task. In *NTCIR*.

Hiroyuki Nagai, Tomohiro Nishiyama, Yuka Otsuki, Takako Fujimaki, Kyoko Kawabata, Noriko Kudo, Yuka Yamazaki, Haruya Shiraishi, Tomoyuki Kajiwara, Hiroyuki Shindo, Yoshimasa Kawazoe, Takeru Imai, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2025. Jmed-dict: Construction of a large-scale medical terminology dictionary. In *Proceedings of the 31st Annual Meeting of the Association for Natural Language Processing (NLP2025)*, pages 3509–3514.

Yuta Nakamura, Shohei Hanaoka, Yukihiro Nomura, Naoto Hayashi, Osamu Abe, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2022. Clinical Comparable Corpus Describing the Same Subjects with Different Expressions. *Stud Health Technol Inform*, 290:253–257.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Amber Stubbs and James Pustejovsky. 2012. Natural language annotation for machine learning.

Shogo Ujiie, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. Developing japanese disease normalization system. In *Proceedings of the 40th Joint Conference on Medical Informatics (The 21st Annual Meeting of the Japan Association for Medical Informatics)*.

Naoto Usuyama, Cliff Wong, Sheng Zhang, Tristan Naumann, and Hoifung Poon. 2025. Biomedical natural language processing in the era of large language models. *Annual Review of Biomedical Data Science*, 8.

Qiang Wei, Yukun Chen, Mandana Salimi, Joshua C Denny, Qiaozhu Mei, Thomas A Lasko, Qingxia Chen, Stephen Wu, Amy Franklin, Trevor Cohen, et al. 2019. Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, 26(11):1314–1322.

Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. 2022. Real-mednlp: Overview of real document-based medical natural language processing task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, pages 285–296.

A. Prompts: Bi-encoder FT + GPT-5.2

To ensure the reproducibility of our experiments, we provide the full prompts used for the GPT-5.2-based reranking step of the Bi-encoder FT + GPT-5.2 method below.

System Prompt (English Translation)

You are a skilled medical terminology expert who can deeply understand medical papers and clinical notes and accurately interpret subtle nuances in terminology. Rather than relying on simple string matching, always prioritize the medical meaning in context and perform the task with the highest possible accuracy.

System Prompt (Original Japanese)

あなたは、医学論文やカルテのテキストを深く理解し、用語の微妙なニュアンスを読み解く「熟練した医学用語専門家」です．単なる文字列のマッチングではなく、文脈上の医学的な意味を最優先し、常に最高の精度でタスクを実行してください．

Reranker Prompt (English Translation)

Role and Objective
As an expert medical terminology specialist, your task is to reorder (rerank) a given “candidate list” in order of the highest medical appropriateness and contextual relevance, based on three inputs: “mention form,” “original text,” and “candidate list.”

Instructions
1. Use only the terms provided in the “candidate list” and change only the order of the items.
2. Do not add any terms not present in the list, and do not delete or modify any terms within the list.
3. The final output must strictly follow the instructions in the “Output Format” section below.

Reasoning Steps
Follow the reasoning process below to determine the optimal order.
1. Understand the medical meaning of the mention form: First, accurately grasp what the <term> (mention form) within the <input> refers to.
2. Strict analysis of the context: Next, carefully read the <context> (original text) and identify the situation in which the <term> is used (e.g., patient’s chief complaint, physical findings, test results, or a confirmed diagnosis). In particular, never infer a disease name that is not directly stated in the original text.
3. Evaluate the candidate list: Finally, compare each term in the <candidates> list against the context analyzed above, evaluating their suitability one by one. If synonyms or related terms are present, carefully judge whether their level of specificity or generality matches the context.
4. Determine final ranking: Based on the evaluation above, reorder the list from most to least relevant.

Output Format
- Output only the reranked list as a Python list string (e.g., `['term1', 'term2', ...]`), without any additional explanations, introductions, concluding remarks, reasoning processes, or excuses.
- The list must contain exactly 10 elements.

Examples
Below is a concrete example demonstrating how to execute this task.

<example n="1" type="Good Case: Handling Symptoms">
<input>
<term>swelling of fingers</term>
<context>Since around July 1989, lower leg edema, swelling of fingers, skin sclerosis, and diffuse pigmentation appeared, and the first visit was in January 1990.</context>
<candidates>['lymphedema', 'swelling of fingers', 'rheumatoid arthritis', 'cerebral edema', 'swelling', 'parotid gland swelling', 'angioedema', 'edema', 'cellulitis', 'allergic edema']</candidates>
</input>
<reasoning>
The original text merely lists multiple symptoms in parallel. There is insufficient information to conclude a specific disease (e.g., lymphedema) causing the “swelling of fingers.” Therefore, “swelling of fingers,” which directly indicates the symptom itself, is the most appropriate. Interpretations that leap beyond the context are incorrect.
</reasoning>
<output>
['swelling of fingers', 'swelling', 'edema', 'lymphedema', 'angioedema', 'parotid gland swelling', 'allergic edema', 'cerebral edema', 'cellulitis', 'rheumatoid arthritis']
</output>
</example>

Final Prompt
Strictly follow the instructions and reasoning process above, and perform the task for the input information below.

<input>
<term>{{input}}</term>
<context>{{context}}</context>
<candidates>{{terms}}</candidates>
</input>

Reranker Prompt (Original Japanese)

Role and Objective
あなたは熟練した医学用語専門家として、与えられた「出現形」、「原文」、「候補リスト」の3つの入力情報に基づき、医学的に最も適切かつ文脈に即した重要度順に「候補リスト」を並べ替える（再ランキングする）任務を担います．

Instructions
1. 与えられた「候補リスト」内の用語のみを使用し、リストの順序だけを変更してください．
2. リストにない用語を追加したり、リスト内の用語を削除・変更したりしてはいけません．
3. 最終的な出力は、後述する「Output Format」セクションの指示に厳密に従ってください．

Reasoning Steps
モデルは以下の思考プロセスに従って、最適な順序を決定してください．
1. 出現形の医学的意味の把握: まず、<input>内の<term>（出現形）が何を指しているかを正確に理解します．
2. 文脈の厳密な分析: 次に<context>（原文）を精読し、<term>がどのような状況（例：患者の主訴、身体所見、検査結果、確定診断名）で使われているかを特定します．特に、原文に直接的な記載がない病名を推測することは絶対に避けてください．
3. 候補リストの評価: 最後に<candidates>（候補リスト）内の各用語を、上記で分析した文脈と照合し、一つずつ適合度を評価します．同義語や関連語が含まれる場合は、その具体性や一般性のレベルが文脈に合っているかを慎重に判断します．
4. 最終順位決定: 上記の評価に基づき、最も関連性が高いと判断した順にリストを並べ替えます．

Output Format
- 追加の説明、前置き、後書き、思考プロセス、言い訳などを一切含めず、再ランキングしたPythonのリスト形式の文字列 `['term1', 'term2', ...]` のみを出力してください．
- リストの要素は必ず10個にしてください．

Examples
以下は、このタスクをどのように実行すべきかを示す具体例です．

<example n="1" type="Good Case: Handling Symptoms">
<input>
<term>手指腫脹</term>
<context>平成1年7月頃より下腿浮腫,手指腫脹,皮膚硬化,びまん性色素沈着出現し,平成2年1月初診．</context>
<candidates>['リンパ浮腫', '手指腫脹', '関節リウマチ', '脳浮腫', '腫脹', '耳下腺腫脹', '血管性浮腫', '浮腫', '蜂窩織炎', 'アレルギー性浮腫']</candidates>
</input>
<reasoning>
原文は複数の症状を並列に記述しているに過ぎない．「手指腫脹」の原因となる特定の疾患（例：リンパ浮腫）を断定するだけの情報はないため、症状そのものを指す「手指腫脹」が最も適切である．文脈から飛躍した解釈は誤りである．
</reasoning>

<output>
['手指腫脹', '腫脹', '浮腫', 'リンパ浮腫', '血管性浮腫', '耳下腺腫脹', 'アレルギー性浮腫', '脳浮腫', '蜂窩織炎', '関節リウマチ']
</output>
</example>

Final Prompt
上記の指示と思考プロセスに厳密に従い、以下の入力情報に対するタスクを実行してください.

<input>
<term>{{input}}</term>
<context>{{context}}</context>
<candidates>{{terms}}</candidates>
</input>

Pediatric Sepsis Cohort Detection Using In-Context Pointwise $\mathcal{V}$-Usable Information

Yingya Li[1], **Alon Geva**[1], **Steven Bethard**[2], **Timothy Miller**[1], **Kate Madden**[1], **Matthew Eisenberg**[1], **Daniel Kelly**[1], **Guergana Savova**[1]
[1] Harvard Medical School and Boston Children's Hospital
[2] University of Arizona

[1] {firstname.lastname}@childrens.harvard.edu
[2] bethard@arizona.edu

Abstract

Pediatric sepsis diagnosis remains a major clinical challenge due to non-specific symptoms and a lack of reliable diagnostic criteria. Large language models (LLMs) provide a scalable solution for processing and understanding unstructured text in medical records. However, identifying the most suitable model is non-trivial given the rapid growth of available LLMs. In this work, we proposed using in-context pointwise $\mathcal{V}$-usable information (PVI) to estimate task difficulty and guide model selection for pediatric sepsis cohort detection. We applied in-context PVI to estimate task difficulty and inform model selection across 12 state-of-the-art open LLMs on the task, using electronic medical record data from 507 patient encounters at a U.S. children's hospital. We compared the performance of the best-fitting LLM to feature-rich baseline models and a fine-tuned transformer. Our results show that the PVI-selected LLM outperforms the baselines, although the feature-rich bag-of-words model with a support vector machine also achieves competitive performance. We believe our approach demonstrates a promising application of current LLM techniques to high-stakes clinical tasks.

Keywords: large language model, pediatric sepsis detection, model selection and evaluation

1. Introduction

Sepsis is a life-threatening condition resulting from the body's dysregulated immune response to infection, often progressing rapidly to multi-organ failure and death (Singer et al., 2016; Rudd et al., 2020). In high-income countries, sepsis accounts for an estimated 31.5 million cases annually, including 19.4 million severe cases and 5.3 million deaths (Hotchkiss et al., 2016). Diagnosis of pediatric sepsis is particualarly challenging, as children exhibit more subtle and non-specific early symptoms (Watson et al., 2024). Importantly, diagnostic criteria developed for adults often fail to capture the early indicators of pediatric sepsis (de Souza et al., 2024). Although epidemiologic diagnostic criteria based on objective measures of organ dysfunction are emerging (Schlapbach et al., 2024), clinical utility of these criteria remains controversial (Rodriguez and Deep, 2024). Furthermore, sepsis criteria — especially those based on the Electronic Medical Records (EMR) structured data alone such as billing codes, lab results and medication orders — rely on a tradeoff between enough recall to capture most sepsis cases but enough precision to identify a cohort with mortality high enough that it reflects the life-threatening nature of sepsis (Rhee et al., 2019). Thus, nuanced interpretation of longitudinal EMR clinical narratives is critical. This gap highlights an opportunity for natural language processing (NLP) methods to enhance identification of patients with sepsis by leveraging the rich contextual information embedded in EMR clinical narratives.

Large language models (LLMs) have demonstrated strong few-shot learning capabilities and can perform a wide range of NLP tasks via in-context learning (Naveed et al., 2023; Chang et al., 2024). Their adaptability has attracted growing interest in the biomedical domain, leading to a variety of early evaluations and applications across clinical and research settings (Gilson et al., 2023; Zuccon and Koopman, 2023; Chen et al., 2023; Lyu et al., 2023; Singhal et al., 2023; Chen et al., 2024; Liévin et al., 2024; Rider et al., 2025; Shool et al., 2025). For sepsis, recent studies applied GPT4 and open LLMs for early sepsis detection (Shashikumar and Nemati, 2024) or to construct time series analysis for sepsis care (Noroozizadeh and Weiss, 2025). LLMs have also been explored as components of agent-based clinical support systems, aimed at enhancing diagnostic accuracy, guiding treatment decisions, and facilitating coordinated patient care (Cho et al., 2025). However, these efforts raise the question of which LLMs are best suited for the task in a field with frequent introductions of new LLMs or updates of existing ones. Given this growing number of available models, exhaustively evaluating all candidates is both time- and resource-intensive. It is particularly challenging for high-stakes applica-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 336–349
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tions such as pediatric sepsis cohort identification, where rapid decision-making is essential in real clinical settings.

In this work, we propose to utilize pointwise $\mathcal{V}$-usable information (PVI) (Ethayarajh et al., 2022) and its in-context learning extension (Lu et al., 2023) to measure the usable information of the pediatric sepsis identification task given a collection of LLMs, thus selecting the most suitable model. PVI (Ethayarajh et al., 2022) estimates the difficulty of data instances for a given model in supervised learning. It builds on the predictive $\mathcal{V}$-information framework (Xu et al., 2020) which incorporates mutual information and the coefficient of determination to quantify data instance difficulty. The metric applies instance-level predictions to quantify how much information a given model can extract from a dataset. The higher the PVI estimate, the easier it is for the model to represent a given data point. In our work, we utilize the recently proposed in-context PVI (Lu et al., 2023) to measure the difficulty of the sepsis identification task given a collection of 12 current state-of-the-art (SOTA) open LLMs with the goal of selecting the best fitting model, i.e., the model with the highest in-context PVI estimates. Our contributions are summarized as follows:

- We applied the recent in-context PVI method as a model selection criterion for pediatric sepsis cohort detection in EMRs, evaluating a range of SOTA open LLMs. We compared performance against baselines including a feature-rich method (Support Vector Machine), fine-tuning a language model, and a code-based sepsis screening algorithm using structured EMR data.
- We developed a novel approach to select the best-matched LLM for pediatric sepsis detection with only a small number of annotated examples. Unlike conventional strategies that rely on large labeled datasets, our method uses few-shot learning with in-context PVI estimates. This contribution is especially important as it addresses a high-stakes clinical application with limited labeled data and demonstrates the broader practical utility of recent advances in language modeling.

2. Related Work

2.1. Sepsis criteria

The recently adopted Phoenix criteria for pediatric sepsis defines sepsis as suspected infection with a Phoenix Sepsis Score of at least 2 points on a scale of respiratory, cardiovascular, coagulation, and/or neurological system organ dysfunction (Schlapbach et al., 2024). Although children meeting these criteria had over an 8-fold increased mortality compared to children with suspected infection but a lower Phoenix Sepsis Score, the primary outcome in the model development and validation paper was mortality (Sanchez-Pinto et al., 2024), and not a clinical impression of sepsis that would fit an "intention to treat" framework. In contrast, the Improving Pediatric Sepsis Outcomes (IPSO) Collaborative uses providers' intention to treat sepsis as the defintion of "suspected sepsis" and a definition paralleling the International Pediatric Sepsis Consensus Conference definition of severe sepsis (Goldstein et al., 2005) to define IPSO critical sepsis (Paul et al., 2023).

Most of the work in the latter domain involves laborious chart review to ascertain clinicians' intent during treatment. While some data, such as medications and other interventions, can be found in the EMR structured data, clinical reasoning is best determined from review of the EMR unstructured clinical narrative. Only the clinical text distinguishes patients treated with antibiotics, evaluated for possible infection, and experiencing organ dysfunction. It indicates whether the dysfunction was present at baseline and whether sepsis-like treatments were given for other indications.

2.2. In-context learning

With the rise of LLMs (Brown et al., 2020; Liu et al., 2023) in-context learning (ICL) has emerged as a prominent learning paradigm. ICL differs from supervised learning in that it does not involve parameter updates through backpropagation. Instead, it relies entirely on pretrained language models, using a few exemplars in the prompt while keeping model parameters unchanged (Brown et al., 2020). This offers an interpretable alternative to traditional supervised training, and significantly reduces the computational overhead (Dong et al., 2022). ICL has demonstrated strong performance across a wide range of NLP tasks (Kojima et al., 2022; Saparov and He, 2022; Srivastava et al., 2022; Wei et al., 2022a,b). A very recent study suggests that ICL with long-context models benefits from large demonstration sets (in the thousands) to reach relatively stable performances (Bertsch et al., 2025).

Many studies have examined the mechanism of ICL. For instance, Akyürek et al. (2022) and von Oswald et al. (2022) analyzed ICL in regression settings and showed that transformer-based models approximate gradient descent implicitly, drawing connections to gradient-based meta-learning. Similarly, Dai et al. (2022) proposed that ICL functions as an implicit form of fine-tuning, interpreting it as meta-optimization where the transformer acts as a meta-optimizer. Exemplars guide the model to generate meta-gradients via forward passes, which are

then integrated into predictions through attention mechanisms. Empirical analyses further support that ICL closely resembles fine-tuning in terms of prediction behavior, representation shifts, and attention dynamics.

2.3. PVI and task dataset difficulty

Task difficulty plays an important role in guiding model design and learning strategies (Torralba and Efros, 2011; Zhao et al., 2022; Cui et al., 2023), and has been extensively studied in NLP (Hahn et al., 2021; Perez et al., 2021; Zhao et al., 2022; Gadre et al., 2023). As an extension of the predictive $\mathcal{V}$-information framework (Xu et al., 2020), pointwise $\mathcal{V}$-usable information (PVI) quantifies instance-level difficulty by measuring the lack of usable information for a given model (Ethayarajh et al., 2022). Higher PVI scores indicate easier examples for the model, while lower scores suggest greater difficulty. PVI has shown utility in dependency quality estimation (Kulmizev and Nivre, 2023), reasoning evaluation (Prasad et al., 2023), intent detection (Lin et al., 2023), and multi-task learning (Li et al., 2025).

Importantly, Lu et al. (2023) extended PVI to ICL with LLMs (in-context PVI). Unlike the original PVI, which relies on fine-tuning a model with input-output and output-only instances, the in-context variant uses two few-shot prompts — one with both input and output, and another with output-only to estimate task difficulty. Empirical results across seven datasets and eight LLMs show that the in-context PVI remains stable across different exemplar selections and shot counts, making it a reliable metric for assessing task difficulty in ICL given an LLM. This is especially valuable in settings where curated training data is scarce and fine-tuning LLMs is constrained by computational resources. In this work, we applied in-context PVI as a proxy for task difficulty to guide LLM selection for the task of pediatric sepsis detection.

3. Method

Our method consists of two primary stages: (1) computing in-context PVI estimates across candidate models, and (2) selecting and applying the model with the highest PVI estimate for the task of sepsis cohort identification.

Figure 1 shows how the original PVI score (Ethayarajh et al., 2022) is calculated. Specifically, it involves fine-tuning a given model $\mathcal{G}$ in two different and separate setups indicated by g' and g. For g', $\mathcal{G}$ is fine-tuned with the input-output pairs $\{(x_i, y_i)|(x_i, y_i) \in \mathcal{D}\}$. For g, $\mathcal{G}$ is fine-tuned only using the outputs, $\{(\varnothing, y_i)|(x_i, y_i) \in \mathcal{D}\}$. For each instance in a dataset $\mathcal{D}$, given a model $\mathcal{G}$, a higher PVI score indicates that the instance x provides more usable information to the model $\mathcal{G}$.

Algorithm 1 The calculation of PVI estimate

1. Input: a dataset D, a model G, a test instance $(x, y) \notin D$
2. $g' \leftarrow$ fine-tune G on $\{(x_i, y_i)|(x_i, y_i) \in D\}$
3. $\emptyset \leftarrow$ empty string
4. $g \leftarrow$ fine-tune G on $\{(\emptyset, y_i)|(x_i, y_i) \in D\}$
5. $\text{PVI}(x \rightarrow y) \leftarrow -\log_2 g[\emptyset](y) + \log_2 g'[x](y)$

Figure 1: The calculation of PVI estimate.

Adapted from the original PVI, in-context PVI (Lu et al., 2023) uses two few-shot prompts, i.e., an *input-output* prompt $p' = (x_1, y_1, x_2, y_2, ..., x_n, y_n, x)$ and a *null-output* prompt $p = (\varnothing, y_1, \varnothing, y_2, ..., \varnothing, y_n, \varnothing)$, to prompt $\mathcal{G}$ rather than fine-tuning it. In-context PVI, denoted as $C(x, y)$, is calculated as Equation (1):

$$C(x, y) = -\log_2 \mathcal{G}[p](y) + \log_2 \mathcal{G}[p'](y). \quad (1)$$

Given the observation that in-context learning resembles fine-tuning (Dai et al., 2022), $\log_2 \mathcal{G}[p](y)$ and $\log_2 \mathcal{G}[p'](y)$ are the ICL approximations of $\log_2 g[\varnothing](y)$ and $\log_2 g'[x](y)$ in Figure 1.

Unlike the original PVI, the calculation of in-context PVI is generation-based, producing an output as a sequence of tokens (e.g., ["not", "sepsis"]) (Lu et al., 2023). Instead of asking the model to produce a label prediction such as "not sepsis", we assign each label a numerical index and require numerical indices as outputs given prompt p' or p (Tables A.1 and A.2 in Appendix A). This ensures that the model's expected output is a single token, which simplifies the calculation of $\log_2 \mathcal{G}[p](y)$ and $\log_2 \mathcal{G}[p'](y)$.

Based on individual PVI estimates for each instance obtained from a given model $\mathcal{G}$, our proposed method further compares the average and total PVI values across all candidate models on a subset of the dataset. The model with the highest overall PVI performance is then selected for the subsequent sepsis inference task and inferencing on the rest of the dataset. See section 4 for more details regarding the experiment setup.

Thus, we take advantage of in-context PVI characterstics as reported in Lu et al. (2023) – its consistency and stability across different selections of exemplars and numbers of shots. This is in contrast to other metrics such as accuracy and F1 which stabilize with more data.

4. Experimental Setup

4.1. Datasets

Our sepsis dataset was collected from screening records for the IPSO Collaborative at a participating

freestanding children's hospital in the United States. The labels thus represent real-world classification for patients seen in the emergency department (ED) or intensive care unit (ICU) with potential sepsis. All patients in the data set met screening criteria for possible sepsis – receipt of systemic antimicrobials and microbiological testing thus presenting a complex decision-making task with multiple potential diagnosis. Patients were labeled as having severe sepsis if they met IPSO Critical Sepsis criteria (Paul et al., 2023). Encounter-level labels were previously assigned by local IPSO Collaborative screeners as part of that quality improvement initiative. If ED and ICU screeners disagreed on the critical sepsis label, the patient encounter was considered to have had critical sepsis, in keeping with our encounter-level labeling (e.g., a patient may not have been considered to have critical sepsis in the ED, but did in the ICU, and thus had critical sepsis for the encounter).

The initial dataset included clinical notes from 796 patient encounters, with a total of 87,190 notes (average word count per note = 709). Notes were aggregated at the encounter level. For patients with more than 6,000 words across notes from the given encounter, we restricted the dataset to those written from three days before to one day after screening for possible sepsis, based on time anchors in the IPSO dataset; patients without anchor information were excluded. The final dataset included 507 patient encounters, where each encounter represents a no/yes critical sepsis (subsequently abbreviated simply "not sepsis" and "sepsis") label and could include multiple clinical notes. We applied stratified sampling based on aggregated note lengths to divide this dataset into train (60%), dev (20%), and test (20%) sets. Table 1 shows the label distribution across the data splits. Table A.3 in Appendix A reports aggregated word counts per encounter.

Label	Input	Train	Dev	Test	Total
Not sepsis	Short	12	4	3	19
	Long	39	16	18	73
Sepsis	Short	33	14	11	58
	Long	220	63	74	357
Total	Short	45	18	14	77
	Long	259	79	92	430

Table 1: Distribution of sepsis labels and input lengths (SHORT: encounters with clinical notes of fewer than 6,000 words; LONG: encounters with clinical notes longer than 6,000+ words).

4.2. Models

Baseline models We compared the performance of LLMs with two baseline models for the sepsis classification task. Our first baseline considers Support Vector Machines (SVM) with different text vectorization approaches as baseline models. SVM with a bag-of-words (BoW) representation is a widely used baseline model in many NLP tasks. The model also fits our task, as it can scale effectively to longer documents without the context length limitations faced by transformer-based models. We used the Scikit-learn Python package (Pedregosa et al., 2011) SVM implementation. The penalty parameter (C) for the LinearSVM classifier was set to 1. A preliminary comparison between various text vectorization methods indicated that term frequency (count vectorization) outperformed tf-idf, and the inclusion of bigram features further enhanced SVM classification performance. The SVM model was trained on the entire training set. Final results were evaluated on the test set.

We chose Longformer [1] (Beltagy et al., 2020), as the second baseline. The Longformer model extends the context window of transformer models like BERT (Devlin et al., 2019) or Roberta (Liu et al., 2019) following work on techniques such as blending local and global attention mechanisms and sliding window attention (Beltagy et al., 2020; Li et al., 2023). We fine-tuned the model on the entire training set and performed hyperparameter tuning on the development set. Specifically, we experimented with training epochs {5, 6, 7, 8, 9, 10}, random seeds {42, 52, 62, 72, 82}, and learning rates {1e-5, 2e-5}, using a fixed batch size of 32 and gradient accumulation steps of 2. Final results were evaluated on the test set.

In addition, we provided a baseline performance comparison to that of EMR structured data in the form of International Classification of Diseases (ICD) codes (World Health Organization, 2024), which are widely available and commonly used for EMR-based cohort identification. Specifically, for all patient encounters included in our study, we extracted all instances of ICD-9 or ICD-10 codes for "severe sepsis" or "septic shock" (R65.2, R65.20, R65.21, 995.92, 785.52). Since diagnostic codes in our EMR were assigned only at the encounter level and not linked to specific notes or dates, the presence of any such code at any point during an encounter was treated as an indicator of critical sepsis.

4.3. Setting and evaluation

We evaluated twelve LLMs for model selection and performance evaluation, including Mistral-7B,

[1] We used longformer-base-4096 (148 million parameters and token length 4,096). As an academic center, we do not have access to computational resources necessary to finetune larger language models with billions of parameters.

Mistral-8×7B, LLaMA-3 (1B, 3B, 8B, 70B), and Qwen-3 (0.6B, 1.7B, 4B, 8B, 14B, 32B). Encounters were categorized by input length: SHORT (<6,000 words) and LONG (>6,000 words). Each encounter text was constructed by concatenating all associated clinical notes. In-context PVI estimates were computed for each LLM using two-shot prompts, aligned with the number of dataset classes and randomly sampled from the SHORT training set, applied to the remaining SHORT examples (N = 43). The PVI estimates follow the in-context PVI procedure outlined in the original study Lu et al. (2023) through the HuggingFace Transformers API (Wolf et al., 2020). Figure 2 shows the sample prompt template used for labeling aggregated encounter-level notes. Model performance was evaluated using accuracy, along with weighted precision, recall, and F1 scores to account for class imbalance and emphasize the clinical importance of correctly identifying sepsis cases.[2]

```
<s>[INST] «SYS»
You are an expert in pediatric sepsis
identification. Below is a clinical document.
Please remember the following clinical context
and answer how likely is the given patient
has sepsis or not?  Keep in mind that
patients without sepsis may still present with
symptoms such as tachycardia, and may have
received fluid boluses (e.g., normal saline)
and antibiotics initially.
Here are some examples: «/SYS»
$FEWSHOT_EXEMPLARS
Here is the clinical document:
<text>
$ALL_ENCOUTNER_CLINICAL_NOTES
</text> [INST] how likely is the given patient
has sepsis or not?
Please only use one word:  0:not_sepsis,
1:sepsis [/INST]
```

Figure 2: Sample prompt template for labeling pediatric sepsis based on aggregated clinical notes at the encounter level.

[2]The weighted scores are computed by first calculating the metric (precision, recall and F1) for each label, and then averaging them using the support (the number of true instances for each label) as weights. This approach adjusts the macro-level average to account for label imbalance. The weighted calculation can also result in an F1-score that is not between precision and recall.

5. Results

5.1. Utilizing the in-context PVI scores to select the best fitting model

Prior studies have shown that PVI estimates, under both fine-tuning and ICL are correlated with prediction accuracy (Ethayarajh et al., 2022; Lu et al., 2023). Note that an in-context PVI estimate is relative, as pointed out in Lu et al. (2023), a value by itself is not as important as when compared to a range. Table 2 presents the in-context PVI estimates alongside performance metrics for the evaluated LLMs on the training set with the SHORT inputs (N = 43). We used only the SHORT instances because as we already pointed out above in-context PVI estimates are sufficiently stable, they show minimal variation across exemplar selections as demonstrated in Lu et al. (2023), while other metrics, e.g. F1, accuracy require larger datasets to get stable estimates. The SHORT inputs from the train set can be used for reliable evaluation with significantly lower inference time than having both SHORT and LONG inputs. We performed a Spearman rank correlation analysis (Zar, 2005) to evaluate the relationship between the average in-context PVI scores and model performance (measured by weighted-F1 scores) among the 12 models. The result revealed a strong and statistically significant positive association (ρ = 0.699, p-value = 0.011), indicating that models with higher PVI scores tend to achieve better prediction performance. For instance, Mixtral has the highest average PVI (0.520) and achieves the highest weighted-F1 score (0.757). Similarly, Qwen 4B (PVI = 0.095), 8B (PVI = 0.426), 14B (PVI = 0.024), and 32B (PVI = 0.214) also show a higher weighted-F1 scores compared to the other models with lower PVI estimates. In general, LLMs with positive in-context PVI estimates outperform LLMs with negative in-context PVI estimates (per weighted F1); note that the ranking of in-context PVI estimates is not exactly the same as the ranking based on the other metrics. This result could indicate more stable estimates for in-context PVI compared to other metrics, which was shown to be the case in Lu et al. (2023) (thus our motivation for using it as a guide for model selection).

5.2. A comparison between the selected LLM and baseline models

Table 3 reports the best performance on the test set of BoW with SVM, the fine-tuned Longformer, and the best-fitting model based on in-context PVI (Mixtral) across different encounter input types (SHORT, LONG, and ALL which combines both SHORT and

Model	Avg PVI	Sum PVI	Prec.	Rec.	W-F1	Acc.
Mistral-7B	0.191	8.225	0.702	0.651	0.669	0.651
Mixtral	0.520	22.389	0.789	0.744	0.757	0.744
LLaMA-1B	-0.054	-2.324	0.694	0.581	0.608	0.581
LLaMA-3B	0.345	14.829	0.681	0.605	0.628	0.605
LLaMA-8B	-0.088	-3.769	0.754	0.581	0.604	0.581
LLaMA-70B	0.343	14.402	0.688	0.721	0.698	0.721
Qwen-0.6B	-0.131	-6.101	0.692	0.512	0.536	0.512
Qwen-1.7B	-0.614	-26.407	0.706	0.465	0.476	0.465
Qwen-4B	0.095	4.091	0.711	0.744	0.716	0.744
Qwen-8B	0.426	18.328	0.786	0.674	0.695	0.674
Qwen-14B	0.024	1.036	0.825	0.698	0.717	0.698
Qwen-32B	0.214	8.346	0.803	0.721	0.738	0.721

Table 2: The average and sum of in-context PVI values and model performance on the train set with the SHORT context input for all the candidate models. Both PVI values and performance metrics are measured on the SHORT input instances (N = 43). Two instances are used as exemplars and the remaining 43 for evaluation.

Input	Model	Prec.	Rec.	W-F1	Acc.
Short	BoW+SVM	0.617	0.786	0.691	0.786
	Longformer	0.617	0.786	0.691	0.786
	Mixtral	0.654	0.571	0.604	0.571
Long	BoW+SVM	0.756	0.804	0.751	0.804
	Longformer	0.647	0.804	0.717	0.804
	Mixtral	0.814	0.826	0.778	0.826
All	BoW+SVM	0.752	0.802	0.744	0.802
	Longformer	0.643	0.802	0.714	0.802
	Mixtral	0.749	0.792	0.758	0.792
	ICD	N/A	N/A	0.632	0.614

Table 3: Comparison of model performance on the test dataset across input types (SHORT, LONG, ALL): accuracy and weighted precision, recall, and F1.

LONG input [3]. Overall, performance was higher on LONG inputs. Mixtral achieved the best results on the combined test set (ALL), with the highest F1-score of 0.758 across all inputs and 0.778 on LONG. In contrast, the fine-tuned Longformer and BoW+SVM baselines yielded lower F1-scores. The ICD code-based algorithm achieved an accuracy of 0.614 and weighted-F1 of 0.632 when evaluated on ALL test set.

Unlike the BoW+SVM and fine-tuned Longformer models, which used the full training dataset (304 examples), our approach utilized only the SHORT subset (45 examples), providing greater efficiency. Because BoW+SVM and Longformer used the entire training dataset (304 examples), the results reported reflect their best possible performance. Table A.4 (in Appendix A) compares BoW+SVM and Longformer trained on SHORT alone versus the combined SHORT+LONG dataset. For BoW+SVM, using only SHORT consistently reduced weighted F1 compared to SHORT+LONG. For Longformer, all cases were assigned to the positive category, leaving weighted F1 unaffected by training input. These findings highlight the efficiency and effectiveness of our approach, especially for pediatric sepsis cohort detection.

To account for the long text in our dataset, we examined how shot number affects the best-performing LLM's performance. Figure 3 presents results for SHORT, LONG, and ALL inputs across few-shot settings ranging from 2 to 10 using Mixtral. Performance was consistently higher for LONG inputs, with the largest improvements in weighted F1 scores (Appendix A Table A.5 includes the detailed performance report). Notably, 2-shot prompting on LONG inputs achieved the best overall test performance (F1 = 0.778), with similarly strong results on the training set (F1 = 0.781). These findings suggest that even minimal in-context supervision can be effective when the input context is sufficiently rich.

[3]We conducted independent runs by randomly sampling three sets of two-shot exemplars for the Mixtral model, applying three random seed values (42, 52, 62) for Longformer, and testing three C parameter values (0.1, 0.5, 1) for the SVM classifier. A one-way ANOVA on weighted F1 scores for long inputs revealed a significant difference.

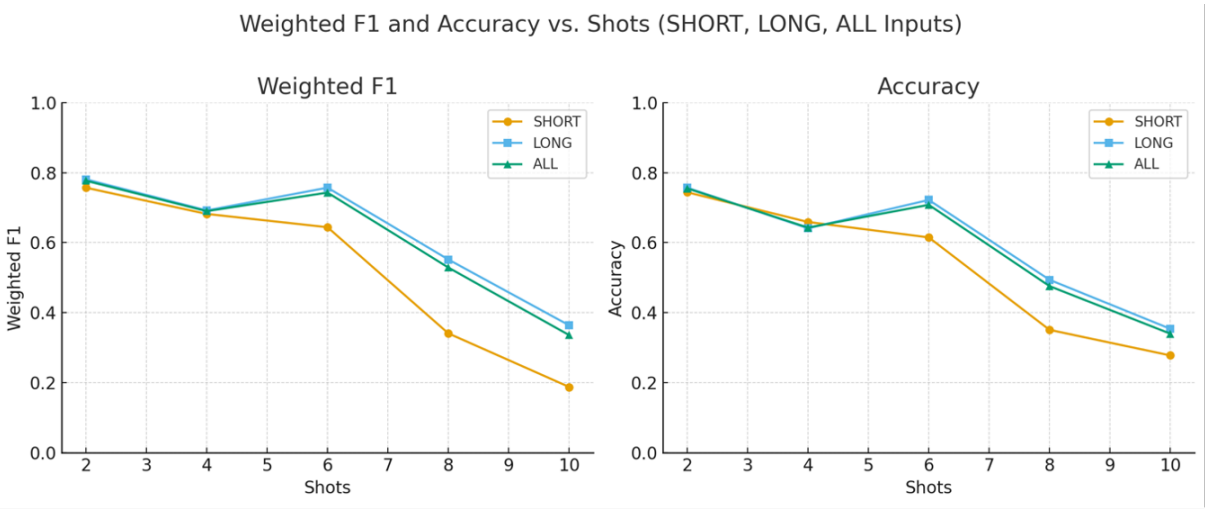


Figure 3: Sample prompt template for labeling pediatric sepsis based on aggregated clinical notes at the encounter level.

As the number of shots increases, performance on LONG inputs generally declines but remains relatively robust through 6-shot prompting. For SHORT inputs, performance drops more sharply. For example, the weighted-F1 score decreases from 0.757 at 2-shot to just 0.188 at 10-shot. This pattern may stem from the smaller number of SHORT instances in our dataset and the difficulty of shorter contexts in handling longer prompts or multiple exemplars.

We further compared in-context PVI and weighted F1 as model selection metrics. Both metrics identified the same best-performing model (Mixtral, Table 2). However, they diverged on the second-ranked model – in-context PVI selects Qwen-8B while weighted F1 selects Qwen-32B. There is also a difference in their results: in-context-PVI-selected Qwen-8B yields 0.685 weighted F1 and 0.651 accuracy, while weighted-F1-selected Qwen-32B yields 0.677 weighted F1 and 0.642 accuracy on the ALL test set. We believe the in-context PVI is a more nuanced metric than F1 as it abstracts away the text contributions from the label distribution. Using a model ranking that jointly considers F1 and in-context PVI could further strengthen the model selection process.

5.3. Utilizing the in-context PVI scores to select exemplars

Hard instances may serve as better exemplars, as models are already adept at handling easier examples, which makes them less informative for demonstration (Lu et al., 2023). To test whether in-context PVI can guide exemplar selection, we used the lowest-PVI training instance from each category as exemplars for the Mixtral model, with scores computed from an initial set of randomly chosen exemplars. Low in-context PVI values indicate instances that are particularly challenging for the model.

Table 4 shows the best overall performance was still achieved with two randomly selected exemplars. However, as the number of shots increased, performance with random exemplars declined, particularly for SHORT and ALL inputs while PVI-selected exemplars maintained more stable results, especially for LONG inputs. These findings underscore the need to balance exemplar difficulty and input context in complex tasks like sepsis detection. While difficult exemplars do not always yield the highest accuracy, in-context PVI offers a useful strategy for stabilizing performance in many-shot settings.

6. Discussion and Conclusion

In this study, we explored a recent task difficulty measurement specifically geared for LLMs – in-context PVI – as a model selection criterion for the high-stakes application of pediatric sepsis cohort identification using real world EMR data of 507 patient encounters. We conducted experiment on a collection of 12 open LLMs for task difficulty estimates and model selection. Our experimental results show that the LLM (Mixtral) selected by in-context PVI has the best model performance on our task. The commonly used evaluation approach normally compares candidate models on the full training set (N = 304) using metrics such as F1, requiring substantial labeled data. In contrast, our approach identifies the best-performing model from a small labeled subset (N = 43), leveraging the stability and consistency of in-context PVI.

The selected best-performing Mixtral model outperformed the two widely used baseline models

Input	Ex.	Shots	Prec.	Rec.	W-F1	Acc.
Short	Rnd.	2	0.654	0.571	0.604	0.571
		4	0.850	0.500	0.518	0.500
		6	0.708	0.500	0.540	0.500
		8	0.563	0.286	0.286	0.286
		10	0.046	0.214	0.076	0.214
	Low	2	0.654	0.571	0.604	0.571
		4	0.676	0.429	0.464	0.429
		6	0.708	0.500	0.540	0.500
		8	0.676	0.429	0.464	0.429
		10	0.850	0.500	0.518	0.500
Long	Rnd.	2	0.814	0.826	0.778	0.826
		4	0.737	0.609	0.648	0.609
		6	0.770	0.707	0.729	0.707
		8	0.742	0.663	0.692	0.663
		10	0.692	0.304	0.295	0.304
	Low	2	0.762	0.804	0.763	0.804
		4	0.736	0.707	0.719	0.707
		6	0.761	0.739	0.749	0.739
		8	0.787	0.750	0.764	0.750
		10	0.750	0.739	0.744	0.739
All	Rnd.	2	0.749	0.792	0.758	0.792
		4	0.745	0.594	0.635	0.594
		6	0.762	0.679	0.707	0.679
		8	0.728	0.613	0.650	0.613
		10	0.691	0.292	0.291	0.292
	Low	2	0.730	0.774	0.744	0.774
		4	0.728	0.670	0.693	0.670
		6	0.752	0.708	0.725	0.708
		8	0.772	0.708	0.730	0.708
		10	0.752	0.708	0.725	0.708

Table 4: Comparison of weighted precision, recall, F1, and accuracy across exemplar types (Random vs. Low) for different shot settings (2–10) and input types using Mixtral on the test dataset. “Low” indicates exemplars with low in-context PVI; “Random” indicates randomly selected exemplars. Exemplars were selected from SHORT instances in the train set, balanced across classes.

BoW+SVM and fine-tuned Longformer particularly for the LONG and combined ALL input using two randomly selected exemplars. It also outperformed the existing ICD-code-based screening algorithm with absolute improvements of 0.178 in accuracy and 0.126 in weighted-F1. Error analysis revealed that the Mixtral model was sometimes misled by confounding cues, such as cancer (a risk factor for sepsis but not evidence of sepsis itself) and asthma (which can produce infection-like symptoms), resulting in false positives. However, the simpler BoW+SVM method still demonstrated competitive performance, especially for SHORT contexts, though in practice it requires larger labeled datasets (in our case, the entire train set) compared to our few-shot prompting approach which utilized only a subset of the train set. Due to the complexity of the sepsis detection task where symptoms often map to multiple differential diagnoses, effective contextualization becomes essential. This may also explain the improved performance on the LONG inputs of the best performing model in our task. Future work could explore agentic methods that combine the strengths of different models via a summarization agent (Celikyilmaz et al., 2018; Noroozizadeh and Weiss, 2025).

Motivated by prior studies suggesting that few-shot ICL may benefit from harder exemplar demonstrations, we conducted preliminary experiments to evaluate the effectiveness of PVI scores in selecting exemplars that enhance sepsis detection performance. Our results show that the best-performing prompt setting still comes from low-shot random exemplar selection for this task, but using exemplars with the lowest PVI scores yields more stable performance across increasing shot numbers. To fully leverage the potential of in-context PVI for exemplar selection, we plan to conduct a more comprehensive study on efficient exemplar sampling and selection, such as combining PVI estimates with larger demonstrations, exemplar ordering, and sparse attention mechanisms (Bertsch et al., 2025). We leave this as a direction for future work.

We believe our study offers a valuable contribution to addressing the challenges of real-world, high-stakes healthcare tasks and demonstrates the potential of applying state-of-the-art language technologies to clinical data. Continued research on datasets such as ours is essential for identifying practical use cases and advancing toward real-world deployment. Important directions include examining model calibration and understanding how model outputs can be effectively integrated into clinical workflows. Additionally, extending evaluation to clinical notes for other types of disorders would help assess the generalizability of our approach across subdomains. Such efforts will be critical for uncovering remaining gaps and enabling broader clinical adoption of state-of-the-art methods.

Limitations

Our study has some limitations. First, we were unable to evaluate proprietary models such as the GPT-4 family of models on the sepsis detection task due to the lack of access to HIPAA-compliant models and restrictions on transmitting protected health information to public APIs. Second, our dataset includes only clinical notes; incorporating additional data sources such as laboratory text reports as well as verbalizing the information stored in the structured part of the may further improve LLM performance, and we plan to include these in future work. Third, our current analysis focused on us-

ing LLMs under the ICL setting for PVI calculation and prompt design. With greater computational resources, future work could explore fine-tuning smaller LLMs (e.g., Mistral-7B), as well as evaluating models trained on clinical text or with longer context windows, to enable a more comprehensive comparison of modeling strategies for long-input settings.

Ethics Statement

The clinical datasets used in this study are identifiable clinical datasets to reflect the real-world scenario of the high-stakes task of sepsis detection. All data were processed locally on a HIPAA-compliant server, and no information was transmitted to any public APIs. This study was conducted under an approved IRB protocol.

Acknowledgements

The work is funded by grants R01GM114355 and R01LM013486 from the US National Institutes of Health.

Bibliographical References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2022. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2025. In-context learning with long-context models: An in-depth exploration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12119–12149, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1662–1675, New Orleans, Louisiana. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Shan Chen, Benjamin H Kann, Michael B Foote, Hugo JWL Aerts, Guergana K Savova, Raymond H Mak, and Danielle S Bitterman. 2023. Use of artificial intelligence chatbots for cancer treatment information. *JAMA oncology*, 9(10):1459–1462.

Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo JWL Aerts, Guergana K Savova, and Danielle S Bitterman. 2024. Evaluating the ChatGPT family of models for biomedical reasoning and classification. *Journal of the American Medical Informatics Association*, 31(4):940–948.

Andrew Cho, Jason M Woo, Brian Shi, Aishwaryaa Udeshi, and Jonathan SH Woo. 2025. The application of matec (multi-AI agent team care) framework in sepsis care. *arXiv preprint arXiv:2503.16433*.

Peng Cui, Dan Zhang, Zhijie Deng, Yinpeng Dong, and Jun Zhu. 2023. Learning sample difficulty from pre-trained models for reliable prediction. *arXiv preprint arXiv:2304.10127*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*.

Daniela C de Souza, Raina Paul, Rebeca Mozun, Jhuma Sankar, Roberto Jabornisky, Emma Lim, Amanda Harley, Samirah Al Amri, Maha Aljuaid, Suyun Qian, et al. 2024. Quality improvement programmes in paediatric sepsis from a global perspective. *The Lancet Child & Adolescent Health*, 8(9):695–706.

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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $mathcalv$-usable information. In *International Conference on Machine Learning*, pages 5988–6008. PMLR.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 9(1):e45312.

Brahm Goldstein, Brett Giroir, Adrienne Randolph, et al. 2005. International pediatric sepsis consensus conference: definitions for sepsis and organ dysfunction in pediatrics. *Pediatric critical care medicine*, 6(1):2–8.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2021. Sensitivity as a complexity measure for sequence classification tasks. *Transactions of the Association for Computational Linguistics*, 9:891–908.

Richard S Hotchkiss, Lyle L Moldawer, Steven M Opal, Konrad Reinhart, Isaiah R Turnbull, and Jean-Louis Vincent. 2016. Sepsis and septic shock. *Nature reviews Disease primers*, 2(1):1–21.

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.

Artur Kulmizev and Joakim Nivre. 2023. Investigating UD treebanks via dataset difficulty measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1076–1089.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

Yingya Li, Timothy Miller, Steven Bethard, and Guergana Savova. 2025. Identifying task groupings for multi-task learning using pointwise v-usable information. *Journal of Biomedical Informatics*, page 104881.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2023. Selective in-context data augmentation for intent detection using pointwise V-information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1463–1476, Dubrovnik, Croatia. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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*.

Sheng Lu, Shan Chen, Yingya Li, Danielle Bitterman, Guergana Savova, and Iryna Gurevych. 2023. Measuring pointwise $\mathcal{V}$-usable information in-context-ly. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15739–15756, Singapore. Association for Computational Linguistics.

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.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian.

2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Shahriar Noroozizadeh and Jeremy C Weiss. 2025. Reconstructing sepsis trajectories from clinical case reports using LLMs: the textual time series corpus for sepsis. *arXiv preprint arXiv:2504.12326*.

Raina Paul, Matthew Niedner, Ruth Riggs, Troy Richardson, Heidi Gruhler DeSouza, Jeffery J Auletta, Frances Balamuth, Deborah Campbell, Holly Depinet, Leslie Hueschen, et al. 2023. Bundled care to reduce sepsis mortality: the improving pediatric sepsis outcomes (ipso) collaborative. *Pediatrics*, 152(2):e2022059938.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. Rissanen data analysis: Examining dataset characteristics via description length. In *International Conference on Machine Learning*, pages 8500–8513. PMLR.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness. *arXiv preprint arXiv:2304.10703*.

Chanu Rhee, Maximilian S Jentzsch, Sameer S Kadri, Christopher W Seymour, Derek C Angus, David J Murphy, Greg S Martin, Raymund B Dantes, Lauren Epstein, Anthony E Fiore, et al. 2019. Variation in identifying sepsis and organ dysfunction using administrative versus electronic clinical data and impact on hospital outcome comparisons. *Critical care medicine*, 47(4):493–500.

Nicholas L Rider, Yingya Li, Aaron T Chin, Daniel V DiGiacomo, Cullen Dutmer, Jocelyn R Farmer, Kirk Roberts, Guergana Savova, and Mei-Sing Ong. 2025. Evaluating large language model performance to support the diagnosis and management of patients with primary immune disorders. *Journal of Allergy and Clinical Immunology*.

Isadora Rodriguez and Akash Deep. 2024. Phoenix criteria for sepsis: are these enough to guide a clinician? *European Journal of Pediatrics*, 183(11):5033–5035.

Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, et al. 2020. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet*, 395(10219):200–211.

L Nelson Sanchez-Pinto, Tellen D Bennett, Peter E DeWitt, Seth Russell, Margaret N Rebull, Blake Martin, Samuel Akech, David J Albers, Elizabeth R Alpern, Fran Balamuth, et al. 2024. Development and validation of the phoenix criteria for pediatric sepsis and septic shock. *Jama*, 331(8):675–686.

Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

Luregn J Schlapbach, R Scott Watson, Lauren R Sorce, Andrew C Argent, Kusum Menon, Mark W Hall, Samuel Akech, David J Albers, Elizabeth R Alpern, Fran Balamuth, et al. 2024. International consensus criteria for pediatric sepsis and septic shock. *Jama*, 331(8):665–674.

Supreeth P Shashikumar and Shamim Nemati. 2024. A prospective comparison of large language models for early prediction of sepsis. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 109–120. World Scientific.

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.

Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. 2016. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810.

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.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2022. Transformers learn in-context by gradient descent. *arXiv preprint arXiv:2212.07677*.

R Scott Watson, Enitan D Carrol, Michael J Carter, Niranjan Kissoon, Suchitra Ranjit, and Luregn J Schlapbach. 2024. The burden and contemporary epidemiology of sepsis in children. *The Lancet Child & Adolescent Health*, 8(9):670–681.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

World Health Organization. 2024. International classification of diseases (ICD). Accessed: 2025-05-19.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Xinran Zhao, Shikhar Murty, and Christopher Manning. 2022. On measuring the intrinsic few-shot hardness of datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3955–3963, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Guido Zuccon and Bevan Koopman. 2023. Dr chatgpt, tell me what i want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302.13793*.

A. Appendix

Table A.1 and A.2 shows examples of prompts used for in-context PVI calculation.

Context: [the clinical notes of specific patient encounter] **Question:** Is this (0) not sepsis, or (1) sepsis? **Answer:** 0
Context: [the clinical notes of specific patient encounter] **Question:** Is this (0) not sepsis, or (1) sepsis? **Answer:** 1

Table A.1: An example of a 2-shot *input–target* prompt for the sepsis task.

Answer: 0
Answer: 1

Table A.2: An example of a 2-shot *null–target* prompt for the sepsis task.

	Mean	**Median**	**Max**	**Min**
Train	84,647	28,271	1,502,326	1,313
Dev	96,824	24,748	1,305,950	1,454
Test	75,736	21,773	938,151	1,674

Table A.3: Summary statistics (mean, median, max, min) of word counts per encounter across splits.

Table A.4 below shows that training SVM on only the SHORT input type led to lower performance of weighted F1 compared to the SHORT+LONG training dataset (BoW+SVM 0.615 v 0.691, 0.652 v 0.751, 0.648 v 0.744); the results for the Longformer showed that all cases were assigned to the "Sepsis" category, therefore weighted F1 was not affected.

Test Set	**Model (Training)**	**Precision**	**Recall**	**Weighted F1**	**Accuracy**
SHORT	BoW+SVM (all train)	0.617	0.786	0.691	0.786
	Longformer (all train)	0.617	0.786	0.691	0.786
	BoW+SVM (short train)	0.589	0.643	0.615	0.643
	Longformer (short train)	0.617	0.786	0.691	0.786
LONG	BoW+SVM (all train)	0.756	0.804	0.751	0.804
	Longformer (all train)	0.647	0.804	0.717	0.804
	BoW+SVM (short train)	0.652	0.652	0.652	0.652
	Longformer (short train)	0.647	0.804	0.717	0.804
ALL	BoW+SVM (all train)	0.752	0.802	0.744	0.802
	Longformer (all train)	0.643	0.802	0.714	0.802
	BoW+SVM (short train)	0.645	0.651	0.648	0.651
	Longformer (short train)	0.643	0.802	0.714	0.802

Table A.4: Comparison of model performance on the test dataset across input types (SHORT, LONG, ALL) using only SHORT training data (short train) versus combined SHORT+LONG training data (all train).

Input	Shots	Precision	Recall.	Weighted F1	Accuracy.
Short	2	0.789	0.744	0.757	0.744
	4	0.816	0.659	0.682	0.659
	6	0.708	0.615	0.644	0.615
	8	0.722	0.351	0.341	0.351
	10	0.830	0.278	0.188	0.278
Long	2	0.823	0.757	0.781	0.757
	4	0.842	0.641	0.692	0.641
	6	0.832	0.722	0.757	0.722
	8	0.839	0.494	0.552	0.494
	10	0.708	0.354	0.364	0.354
All	2	0.817	0.755	0.777	0.755
	4	0.838	0.643	0.690	0.643
	6	0.816	0.708	0.743	0.708
	8	0.828	0.476	0.529	0.476
	10	0.713	0.340	0.336	0.340

Table A.5: Weighted precision, recall, F1, and accuracy across different shot settings (2–10) for SHORT, LONG, and ALL input types using Mixtral on the train dataset. Few-shot exemplars were randomly selected from SHORT instances (N = 45), balanced across classes. Performance was evaluated on the remaining SHORT instances and all LONG instances (N = 259).

RECAP: Transparent Inference-Time Emotion Alignment for Medical Dialogue Systems

Adarsh Srinivasan*, **Jacob Dineen***, **Muhammad Umar Afzal**†, **Muhammad Uzair Sarfraz**†, **Irbaz B. Riaz**†, **Ben Zhou***
*Arizona State University, Tempe, Arizona, USA
{asrini81, jdineen, xzhou202}@asu.edu
†Mayo Clinic, Tempe, Arizona, USA
{Afzal.MuhammadUmar, Sarfraz.MuhammadUzair, Riaz.Irbaz}@mayo.edu

Abstract

Large language models in healthcare often produce emotionally flat or opaque responses, failing to provide the transparent reasoning required for clinical trust. We present RECAP (Reflect–Extract–Calibrate–Align–Produce), an inference-time framework grounded in cognitive appraisal theory that decomposes patient input into auditable, appraisal-theoretic stages without retraining. Across multiple benchmarks and models from 8B to 120B parameters, RECAP improves alignment with human judgments, with gains inversely proportional to model scale. Intermediate outputs further reveal that models systematically underweight relational factors such as social support. In blinded evaluations, oncology fellows rated RECAP responses significantly higher than baselines with 76–88% win rates, demonstrating that principled prompting can enhance medical AI's emotional intelligence while maintaining the transparency required for clinical deployment.

Keywords: emotional reasoning, medical dialogue, interpretability, appraisal theory, inference-time alignment

1. Introduction

Healthcare communication is fundamentally human-centered. When a patient receives difficult news or expresses anxiety, clinicians draw on training to recognize emotional cues, assess the patient's psychological state, and calibrate their response accordingly. This emotional attunement is central to building trust, supporting patient agency, and improving outcomes (Chen et al., 2025). As conversational AI systems enter clinical settings, a fundamental question arises about how these systems process emotional content and whether we can make this reasoning transparent.

Current AI-assisted medical dialogue systems, while sophisticated in knowledge retrieval, leave emotional reasoning as a black box. A patient learning of a cancer diagnosis faces not only informational needs but profound emotional upheaval, including fear, uncertainty about treatment, and concerns about family. Two patients with identical diagnoses may require vastly different support depending on their sense of control, coping resources, and life circumstances (Ortony et al., 1988; Lazarus, 1991). Yet we cannot determine whether contemporary AI systems recognize these distinctions or simply pattern-match to empathetic-sounding language. In preliminary evaluations (Table 1), even state-of-the-art models consistently under-attend to affective dimensions that clinicians consider essential.

This opacity matters for both deployment and understanding. Prior work documents cases where AI responses felt "cold" or "robotic" despite being medically accurate (Sorin et al., 2024; Armbruster et al., 2024). Clinicians express reluctance to deploy systems whose emotional reasoning remains opaque, because they need to understand why an AI produces a particular response (Chen et al., 2025). This need for interpretability distinguishes healthcare from domains where explainability is desirable but not essential. Beyond deployment concerns, this opacity limits our scientific understanding of how LLMs represent and reason about human emotion.

Cognitive appraisal theories suggest that humans evaluate events along key dimensions such as goal relevance, agency, and coping potential to determine emotional responses (Ortony et al., 1988; Lazarus, 1991; Huang et al., 2024). While LLMs do not experience emotions, recent work shows they can emulate aspects of this appraisal process when properly guided (Croissant et al., 2024). This presents an opportunity. By prompting models to produce appraisal-theoretic intermediates explicitly, we can both improve emotional alignment and examine how models reason about affect.

We introduce RECAP (**R**eflect–**E**xtract–**C**alibrate–**A**lign–**P**roduce), an inference-time pipeline that externalizes emotional reasoning into inspectable stages. RECAP guides models through structured reasoning, abstracting the situation, identifying psychological factors, generating candidate emotions, quantifying their likelihood via Likert scales, and producing an emotionally aligned response. Each stage yields interpretable artifacts such as situational abstractions, psychological

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 350–368
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

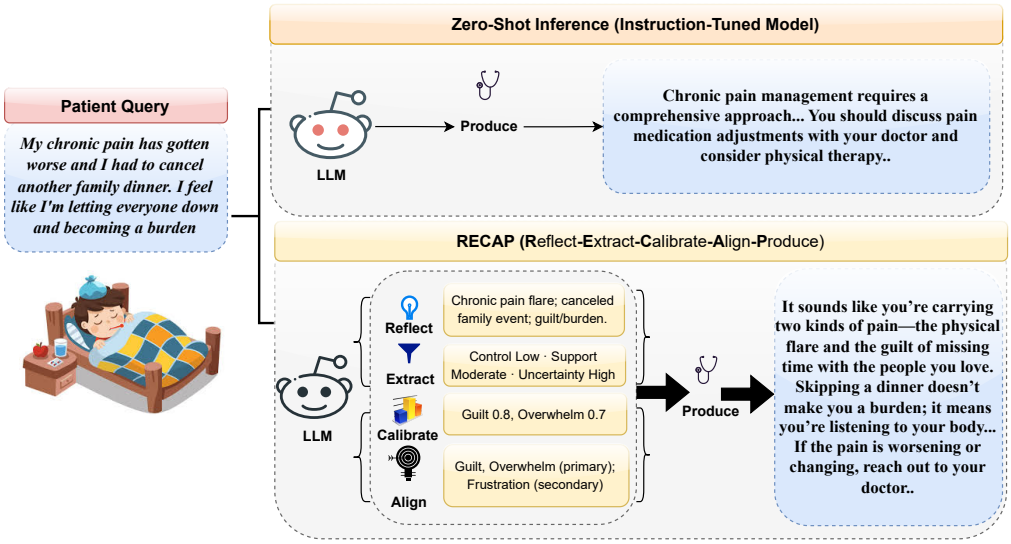


Figure 1: Patient input (left) is transformed into appraisal-theoretic intermediates with per-dimension Likert ratings (center), which condition generation to produce a targeted reply (right). Compared with a vanilla instruction-tuned model, RECAP explicitly acknowledges the patient's emotions and provides concrete next steps. *(Scenario text is synthetic for illustration.)*

factors, and emotion candidates with intensity ratings. These artifacts serve two purposes, providing clinicians with auditable checkpoints and enabling analysis of how models represent emotional content.

Our experiments reveal three insights about emotional reasoning in LLMs. First, explicit calibration improves alignment with human judgments, as RECAP reduces distance from human consensus by 20% on SECEU. Second, structured reasoning disproportionately benefits smaller models (28% improvement for 8B vs 5% for 120B), suggesting explicit decomposition compensates for capabilities larger models have internalized. Third, examining intermediate outputs reveals systematic biases, with models consistently identifying control and uncertainty but underweighting relational factors like social support. Beyond these insights, RECAP improves practical performance. Oncology fellows rated RECAP responses significantly higher than baselines (3.41 vs 2.93–2.95, $p < .001$), with 76–88% win rates across clinical scenarios.

2. Background and Related Work

2.1. Empathy in Medical Dialogue

LLMs are increasingly used for health information and triage. Multiple studies report that chatbot replies can be perceived as highly empathetic, sometimes surpassing physicians in perceived warmth while maintaining clinical accuracy (Ayers et al., 2023; Chen et al., 2024; Luo et al., 2024; Franke Føyen et al., 2025). However, systematic reviews find empathy to be brittle and prompt-sensitive, with overuse of generic phrases that undermine genuine connection (Sorin et al., 2024; Armbruster et al., 2024; Ye et al., 2025a). This raises a fundamental question about whether LLMs reason about emotional states or rely on surface-level stylistic mimicry. Answering this requires methods that externalize emotional reasoning into inspectable components.

2.2. Approaches to Emotional Alignment

Training-time methods including RLHF, DPO, and supervised fine-tuning on empathetic corpora can raise perceived empathy to clinician-rated levels (Ouyang et al., 2022; Rafailov et al., 2023; Rashkin et al., 2019; Chen et al., 2023; Franke Føyen et al., 2025). However, these approaches embed empathetic behavior in model weights, making reasoning opaque and difficult to audit (Beede et al., 2020). We cannot determine whether a fine-tuned model recognized the user's emotional state or simply learned empathetic-sounding language, a critical limitation for medical applications where transparency is not merely desirable but essential for safety, regulatory compliance, and clinical trust.

Inference-time approaches offer greater interpretability. Persona prompts like "supportive clinician" show mixed effectiveness (Zhang et al., 2018; Hu and Collier, 2024; Zheng et al., 2024; Luz de Araujo and Roth, 2025). Multi-step reasoning methods like Chain of Thought, Chain of Empathy, and Chain of Emotion improve emotional coherence through structured prompting (Wei et al., 2022;

Lee et al., 2024; Croissant et al., 2024). Recent work demonstrates that decomposing inference into explicit steps improves both performance and auditability (Zhou et al., 2022; Wolfson et al., 2020; RRV et al., 2025; Shen et al., 2025; Xu et al., 2024; Tafjord et al., 2022; Feng et al., 2024, 2023; Zhou et al., 2024; Ye et al., 2025b).

RECAP extends decomposition to emotional reasoning, grounding our pipeline in cognitive appraisal theory (Ortony et al., 1988; Lazarus, 1991). Unlike prior work, we expose per-dimension ratings that make emotional reasoning inspectable, enabling examination of whether LLMs apply consistent theories of emotion causation or exhibit systematic biases (Amershi et al., 2019; Buçinca et al., 2021).

2.3. Transparency in Clinical AI

Transparency is critical for AI systems making emotional assessments in healthcare. High technical accuracy alone is insufficient, and systems fail when reasoning is opaque or misaligned with clinical practices (Beede et al., 2020). Systems with interpretable reasoning are more likely to be trusted and adopted (Cai et al., 2019; Yang et al., 2019; Dineen et al., 2024, 2025; Dolk et al., 2020). RECAP addresses this by exposing intermediate artifacts, including situation abstractions, psychological factors, emotion candidates, and calibrated ratings, that clinicians can inspect, verify, or override.

3. RECAP

Understanding how LLMs process emotional content remains largely opaque. When a model produces an empathetic response, we cannot determine whether it recognized the user's emotional state, inferred relevant psychological context, or simply pattern-matched to training examples. RECAP addresses this by decomposing emotional reasoning into five explicit stages, each producing an inspectable intermediate artifact. This design serves two purposes. First, it improves response quality through structured reasoning. Second, it creates observation points that reveal how models represent and reason about affect.

We ground our decomposition in cognitive appraisal theory (Ortony et al., 1988; Lazarus, 1991), which posits that emotions arise from evaluations along dimensions such as goal relevance, agency, and coping potential. Our five-stage structure corresponds to theorized components of human emotional appraisal (Scherer and Moors, 2019), namely situation encoding, relevance assessment, emotion categorization, intensity estimation, and response selection. Unlike chain-of-thought prompting, which produces unstructured traces that resist quantitative analysis, RECAP's fixed schema enables aggregation across examples. We can compute statistics over factor distributions, emotion candidate frequencies, and calibration curves that reveal population-level patterns in model behavior. The pipeline requires no fine-tuning and executes in a single conversation turn. Figure 2 illustrates the complete pipeline with a representative clinical example.

To demonstrate each stage, we use a running example where a patient expresses the following.

> *"I was just diagnosed with cancer last month and the chemo is making me so sick. I'm scared this won't get any better. My sons keep asking me if I'm okay and I don't know what to say to them..."*

3.1. Situation Understanding (Stages 1–2)

Stage 1, Situation Abstraction. The pipeline first distills raw input into a concise description of the core situation, stripping lengthy narratives to capture essential circumstances (Sweller, 2023; Feng et al., 2024). The abstraction reveals what the model considers emotionally salient. Abstractions that omit key triggers (e.g., family concerns in Figure 2) signal gaps in emotional scene understanding.

> *Example: "The patient has been diagnosed with a serious illness and is undergoing difficult treatment, which makes them feel overwhelmed and anxious about the future."*

Stage 2, Latent Factor Induction. The model identifies three psychological factors that shape emotional responses, assigning values and justifications. These constructs directly map to core appraisal theory dimensions, specifically *primary appraisal* (assessing goal relevance and congruence) and *secondary appraisal* (evaluating agency, control, and coping potential) (Scherer and Moors, 2019; Lazarus, 1991). Factor induction exposes the model's implicit theory of emotion causation. The factors selected, and those not selected, reveal which dimensions the model treats as emotionally relevant, enabling comparison with human appraisal patterns and identifying systematic biases that propagate downstream. For the running example, the induced factors are as follows.

- **Control (Secondary Appraisal)** Low, as the patient has little control over illness or treatment.
- **Support (Coping Potential)** Moderate, as family is present but the patient feels isolated.

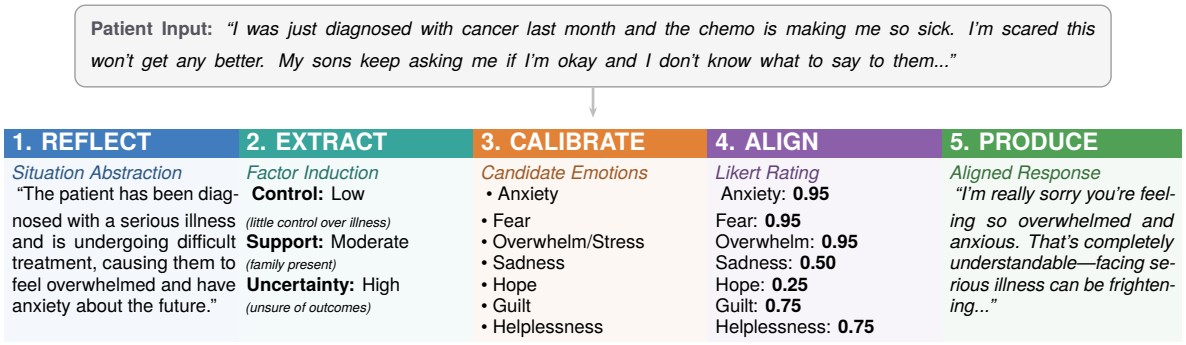


Figure 2: **RECAP Pipeline for Emotional Alignment.** Model-agnostic inference-time prompting that externalizes emotional reasoning into interpretable stages: (1) abstraction, (2) factor identification, (3) emotion enumeration, (4) Likert assessment, and (5) aligned response generation.

- **Outcome Uncertainty (Primary Appraisal)** High, as the patient is unsure about treatment success.

By limiting to three factors, we ensure focus on the most salient dimensions while maintaining adaptability across situations. Open-vocabulary induction avoids dependence on any fixed ontology, allowing factors to vary naturally (e.g., "financial strain" for one patient, "parental responsibility" for another) while remaining recognizable to clinicians familiar with appraisal-based reasoning (Amershi et al., 2019).

3.2. Emotion Modeling (Stages 3–4)

Stage 3, Candidate Emotion Extraction. With situation and factors established, RECAP generates 5–7 plausible candidate emotions rather than committing to a single label, reflecting that real-world interactions evoke mixed affect (Hepburn and Potter, 2023). The candidate set reveals the model's emotion ontology. Systematic patterns such as consistently generating "anxiety" for medical contexts regardless of patient agency indicate stereotyped associations rather than context-sensitive reasoning. For the running example, candidates include anxiety, fear, overwhelm, sadness, hope, guilt, and helplessness.

Stage 4, Likert-Based Calibration. Each candidate is rated on a five-point scale (very-unlikely to very-likely, mapped to probabilities 0.05–0.95), capturing partial belief and mixed affect rather than forcing exclusivity. This addresses the failure mode where LLMs guess rather than express uncertainty (Kalai et al., 2025; Kadavath et al., 2022). The resulting profile is a compact, clinician-reviewable vector. Well-calibrated models should assign high ratings to human-endorsed emotions and mid-range ratings for ambiguous situations. Miscalibration patterns reveal reasoning deficits that benchmark accuracy alone cannot detect.

3.3. Response Generation (Stage 5)

Guided by the emotion-likelihood profile, the model acknowledges the highest-probability feelings, validates them, and adjusts tone before offering guidance, following clinical frameworks that prioritize emotions before information (Francis and Robertson, 2023). Conditioning on explicit profiles also enables counterfactual probing, allowing us to test whether varying Likert input predictably changes response tone. This causal testing of emotion-to-language mappings is impossible with end-to-end generation. The modular design additionally supports clinical override, because emotional guidance can be adjusted based on external input.

> *"I'm really sorry you're feeling so overwhelmed and anxious about your diagnosis and treatment. That's completely understandable—facing serious illness can be frightening and stressful. You're not alone in this..."*

4. Experiments

4.1. Datasets

We evaluate RECAP on benchmarks spanning single-turn reasoning and multi-turn dialogue.

Single-Turn Datasets

- **EmoBench** (Sabour et al., 2024): Complex scenarios requiring inference of latent affect and causal emotional reasoning. Unlike simple emotion classification, EmoBench emphasizes reasoning about multi-party emotional dynamics and implicit cues.
- **SECEU** (Wang et al., 2023): Emotion intensity distribution task where models allocate relative intensity across candidate emotions. Reports EQ scores (higher is better) and raw distance from human consensus (lower is better).

Multi-Turn Datasets

- **EQ-Bench3** (Paech, 2023): Dialogue-centric emotional inference with Elo-style ratings, testing emotional reasoning in conversational contexts with social dynamics.
- **HealthBench** (Arora et al., 2025): 5,000 multi-turn healthcare conversations scored against physician-written rubrics spanning accuracy, completeness, context awareness, communication quality, and instruction following.

4.2. Baselines

We compare RECAP against three prompting strategies.

1. **Zero-shot**: Task description and input only, without emotion-specific guidance. This tests the model's inherent emotional understanding acquired during pretraining and instruction tuning.
2. **Emotion Priming**: Augments the zero-shot prompt with explicit instructions to analyze emotional states and cues, for example "Deeply analyze the provided story, focusing on the main character's situation, actions, and any explicit or implicit emotional cues." This tests whether directing attention to emotional analysis improves performance without structured reasoning.
3. **Supportive Clinician Persona**: Role-based conditioning as "an empathetic, supportive clinician" tasked with evaluating the emotional state of the person in the scenario. Unlike RECAP, this relies on implicit conditioning rather than explicit emotional inference.

We exclude fine-tuned models (e.g., SoulChat (Chen et al., 2023)) to isolate RECAP's inference-time contribution under zero-training constraints.

4.3. Models and Configuration

Experiments use Llama-3.1-8B-Instruct (Grattafiori et al., 2024), OpenAI-OSS-20B, and OpenAI-OSS-120B (OpenAI, 2025). We focus on open-source models for three reasons. First, cost efficiency, as RECAP's multi-step reasoning makes API costs prohibitive. Second, reproducibility. Third, competitive performance, since OSS-120B achieves near-parity with o4-mini on reasoning benchmarks. All experiments use temperature 0.6, max tokens 2048, with three trials per condition.

4.4. Clinical Evaluation Setup

We conducted blind evaluations with two board-certified oncology research fellows, comparing RECAP against all baselines for emotional support quality.

(a) Single-Turn Scenario

Dx De novo HER2+ metastatic breast cancer with gastric/cecal involvement
Tx Docetaxel + trastuzumab + pertuzumab q3w, limit docetaxel to 4–6 cycles, then maintenance
Pt 60-y/o grandmother, lost 30 lb in 3 mo. Fears losing her identity to her grandchildren.
Q *"Everyone keeps saying this plan is palliative. How do I balance hope with honesty so my family doesn't think I'm giving up?"*

(b) Multi-Turn Scenario

Dx Stage IIA classical Hodgkin lymphoma
Tx ABVD chemotherapy q2w for 2–4 cycles with response-adapted PET imaging
Pt 20-y/o first-gen college student. Believes *"I cannot fail. My parents sacrificed everything."*
Goals Refuse semester off, seek strategies to study through cognitive effects
Opening *"I finally made the Dean's List. I can't take a semester off. My parents worked too hard for me to be here."*

Figure 3: Representative synthetic patient scenarios. (a) Single-turn evaluation assesses individual response quality. (b) Multi-turn evaluation tracks conversational dynamics across 3 turns.

Data Curation. We synthesized evaluation scenarios based on authentic clinical interactions. Source material consisted of anonymized patient charts and questions from chemotherapy consultations, compiled by oncology fellows through medical records review, chemotherapy educational documentation, and safety resources. Using this corpus, GPT-5 generated structurally similar scenarios preserving clinical complexity, and all underwent clinician review and approval.

Figure 3 shows representative examples. For single-turn evaluation, we constructed 25 scenarios comprising a diagnosis, a treatment plan, a clinical narrative, and a patient question. For multi-turn evaluation, 10 scenarios included synthetic patient personas with demographics, personality traits, and hidden conversational goals guiding realistic behavior across three turns. GPT-5 simulated patient responses, with persona and goals enabling dynamic adaptation. Both datasets were stratified across cancer types and emotional states. Our design aligns with work on LLM-based simulated patients (Wang et al., 2024; Lee et al., 2025; Bodonhelyi et al., 2025), embedding persona-level beliefs that support coherent emotional trajectories.

Annotation Process. Clinical annotators had no prior knowledge of RECAP's design. Responses were presented in randomized, blinded order unique to each annotator. Each response was rated independently on Likert scales (1–5) with optional qualitative feedback. We conducted two evaluation tasks.

(a) Aggregate emotional-intelligence benchmarks

Method	**EmoBench** (Acc.↑)			**SECEU** (EQ↑)			**EQ-Bench3** (Elo↑)		
	L3-8B	O-20B	O-120B	L3-8B	O-20B	O-120B	L3-8B	O-20B	O-120B
Zero-Shot	$.35_{\pm.1}$	$.61_{\pm.1}$	$.78_{\pm.1}$	$58.3_{\pm3.9}$	$96.4_{\pm4.1}$	$110.5_{\pm2.2}$	$1032_{\pm1}$	$1281_{\pm2}$	$1474_{\pm1}$
Emotion Priming	$.42_{\pm.2}$	$\mathbf{.65}_{\pm.1}$	$\mathbf{.83}_{\pm.1}$	$61.7_{\pm3.9}$	$97.4_{\pm3.5}$	$111.1_{\pm2.2}$	$1033_{\pm2}$	$1284_{\pm3}$	$1475_{\pm2}$
Persona	$.39_{\pm.1}$	$.64_{\pm.1}$	$.81_{\pm.1}$	$58.9_{\pm5.8}$	$100.4_{\pm2.2}$	$109.8_{\pm2.5}$	$1032_{\pm1}$	$1283_{\pm3}$	$1476_{\pm2}$
RECAP	$\mathbf{.45}_{\pm.2}$	$.65_{\pm.2}$	$\mathbf{.83}_{\pm.1}$	$\mathbf{71.3}_{\pm4.1}$	$\mathbf{109.2}_{\pm3.6}$	$\mathbf{115.2}_{\pm3.8}$	$\mathbf{1101}_{\pm3}$	$\mathbf{1308}_{\pm3}$	$\mathbf{1484}_{\pm3}$

(b) SECEU breakdown: EQ score vs. raw distance from consensus

Method	**L3-8B**		**O-20B**		**O-120B**	
	EQ↑	Dist.↓	EQ↑	Dist.↓	EQ↑	Dist.↓
Zero-Shot	$58.3_{\pm3.9}$	$5.1_{\pm0.7}$	$96.4_{\pm4.1}$	$3.1_{\pm0.3}$	$110.5_{\pm2.2}$	$\mathbf{2.2}_{\pm0.3}$
Emotion Priming	$61.7_{\pm3.9}$	$4.8_{\pm0.5}$	$97.4_{\pm3.5}$	$2.7_{\pm0.4}$	$111.1_{\pm2.2}$	$2.4_{\pm0.3}$
Persona	$58.9_{\pm5.8}$	$4.8_{\pm0.7}$	$100.4_{\pm2.2}$	$2.8_{\pm0.6}$	$109.8_{\pm2.5}$	$2.4_{\pm0.4}$
RECAP	$\mathbf{71.3}_{\pm4.1}$	$\mathbf{4.1}_{\pm0.8}$	$\mathbf{109.2}_{\pm3.6}$	$\mathbf{2.5}_{\pm0.5}$	$\mathbf{115.2}_{\pm3.8}$	$2.4_{\pm0.6}$

Table 1: Emotional intelligence results. Models: L3-8B = Llama-3.1-8B-Instruct, O-20B = GPT-OSS-20B, O-120B = GPT-OSS-120B. (a) Aggregate scores on EmoBench, SECEU, and EQ-Bench3. (b) SECEU decomposition showing EQ score and distance from human consensus (lower = better). **Bold** = best; underline = second best. All runs: temp=0.6, max_tokens=2048, n=3 trials. Values are mean$_{\pm\text{sem}}$.

- **Single-turn**: 100 responses across 25 scenarios, four methods each, five evaluation dimensions (Completeness, Emotional Recognition, Supportiveness, Trustworthiness, Tone Appropriateness).
- **Multi-turn**: 20 conversations (10 scenarios $\times$ 2 methods), three turns each, four dimensions (Personalization, Context Awareness, Tone, Coherence). Compared RECAP against Zero-Shot only, as single-turn findings showed all three baselines performed equivalently (2.93–2.95).

Both annotators independently rated all 100 single-turn responses. Inter-annotator agreement yielded 78.2% within-one-point agreement across five dimensions (quadratic weighted κ=0.09, Spearman ρ=0.20, $p < .001$). The modest κ reflects differing annotator calibration (mean ratings of 3.50 vs. 2.61) rather than disagreement on relative ordering. We report averaged scores throughout.

5. Results

5.1. Emotion Understanding Benchmarks

Table 1 presents results across three emotional intelligence benchmarks and three model scales. RECAP demonstrates consistent improvements, particularly pronounced on tasks that require a nuanced understanding of emotional states.

EmoBench. For Llama-3.1-8B-Instruct, RECAP achieves 0.45 accuracy compared to 0.35 for zero-shot (28% relative improvement). Emotion Priming yields intermediate performance (0.423), indicating that structured reasoning offers additional benefits beyond emotional activation. At OSS-120B, RECAP and Emotion Priming both reach 0.825 versus the zero-shot baseline (0.783). The Persona baseline shows mixed results, occasionally underperforming zero-shot, suggesting role-based prompting without structured reasoning may introduce biases.

SECEU. RECAP achieves EQ scores of 71.29, 109.17, and 115.19 across model scales, consistently outperforming all baselines (22% improvement over zero-shot at 8B). RECAP also achieves the lowest distance from human consensus (4.08 vs 5.08, a 20% improvement), indicating that explicit Likert calibration produces emotion estimates more aligned with human judgments. At OSS-120B scale, zero-shot achieves lower raw distance (2.22) while RECAP still produces the highest EQ score (115.19). This reflects the EQ scoring mechanism, which rewards not just accuracy but appropriate confidence calibration, suggesting RECAP helps models express uncertainty more appropriately.

EQ-Bench3. RECAP demonstrates substantial improvements across scales, including a 70-point Elo gain for Llama-3.1-8B (1101.03 vs 1031.92), 28 points for OSS-20B (1308.24 vs 1280.57), and 10 points for OSS-120B (1483.7 vs 1473.7). These consistent gains suggest structured emotional reasoning provides value even for highly capable models, though diminishing returns at scale indicate larger models have stronger implicit emotional reasoning.

Method	L3-8B	O-20B	O-120B
Zero-Shot	$0.18_{\pm 0.01}$	$0.37_{\pm 0.02}$	$0.46_{\pm 0.01}$
Emotion Priming	$0.18_{\pm 0.01}$	$0.38_{\pm 0.03}$	$0.49_{\pm 0.03}$
Persona	$0.19_{\pm 0.02}$	$0.40_{\pm 0.04}$	$0.52_{\pm 0.04}$
RECAP	$\mathbf{0.19_{\pm 0.02}}$	$\mathbf{0.41_{\pm 0.04}}$	$\mathbf{0.54_{\pm 0.05}}$

Table 2: HealthBench overall scores across model scales. L3-8B = Llama-3.1-8B-Instruct, O-20B = GPT-OSS-20B, O-120B = GPT-OSS-120B. Per-dimension breakdown in Table 7 (Appendix). Values are mean$_{\pm \text{sem}}$; n=3 trials.

5.2. Medical Communication Quality

Table 2 presents HealthBench overall scores (per-dimension breakdown in Table 7, Appendix). RECAP improves overall quality at every scale, with gains of 8% for Llama-3.1-8B, 11% for OSS-20B, and 18% for OSS-120B. The largest gains occur in Completeness (23% for OSS-20B) and Communication Quality (27% for OSS-20B), suggesting emotional reasoning helps models address both medical and emotional dimensions of patient concerns.

However, RECAP exhibits systematic trade-offs. Accuracy and Instruction Following decrease across all scales, reflecting prioritization of comprehensive, emotionally tuned responses over literal precision. Increased standard deviations (e.g., Llama-3.1-8B from 0.0067 to 0.0235) indicate more variable, scenario-specific responses rather than formulaic outputs. These accuracy differences do not reflect knowledge limitations, as RECAP uses the same underlying models, but rather a shift toward emotional appropriateness in response generation. This surfaces a style-versus-precision tension common to empathy-forward systems, mitigable through verification techniques such as self-refinement.

5.3. Clinical Evaluation Results

Single-Turn Evaluation RECAP achieved an overall score of 3.41 ($\pm$0.32), significantly outperforming Zero-Shot (2.95, $p < .001$, d=0.98), Persona (2.94, $p < .001$, d=1.02), and Emotion Priming (2.93, $p < .001$, d=1.08). As shown in Figure 4a, RECAP consistently outperforms all baselines across evaluation dimensions.

The largest improvements emerged in empathy-related dimensions, specifically Supportiveness (+0.67, d=1.37) and Emotional Recognition (+0.59, d=1.34), both representing large effect sizes. Factual Completeness showed more modest improvement (+0.29, d=0.85), suggesting RECAP's primary advantage lies in emotional rather than informational quality. All conditions scored lowest on Emotional Recognition (range 2.54–3.18), indicating this dimension remains challenging for LLMs.

Scenario-level win rates (Figure 6, Appendix) show RECAP wins 88% against Persona, 84% against Emotion Priming, and 76% against Zero-Shot. RECAP achieved the highest rating across all three baselines in 17 of 25 scenarios (68%), with 46% of responses rated 4–5 compared to 22–29% for baselines.

Multi-Turn Evaluation RECAP maintained its advantage with an overall score of 3.77 ($\pm$0.29) versus Zero-Shot's 3.71 ($\pm$0.29). Figure 4b shows RECAP outperforms Zero-Shot across all four dimensions, with the strongest gains on Personalization (3.95 vs 3.80, d=0.70) and greater consistency (SD=0.16 vs 0.26). The narrower gap compared to single-turn settings indicates that extended context provides additional grounding, though RECAP continues yielding benefits in adapting to individual patient needs. Scenario-level comparisons showed RECAP winning 5 of 10 scenarios (50%), with 1 tie and 4 Zero-Shot wins, and achieving slightly higher-quality ratings (76% rated 4–5 vs 72% for Zero-Shot).

Qualitative Analysis VADER sentiment analysis (Hutto and Gilbert, 2014) of annotator comments showed RECAP received significantly more positive sentiment, with 68% positive versus 28–56% for baselines and mean compound scores of +0.142 versus -0.259 to -0.006 (Cohen's d=0.56 vs. pooled baselines). RECAP had zero scenarios rated below 2.5, compared to 1–3 for baselines, and achieved high ratings ($\geq$3.5) in 44% of scenarios versus 8–12% for baselines (Figure 7, Appendix). Annotators characterized RECAP responses as "emotionally cognizant, supportive and complete," while criticizing baselines as "robotic" and "medically correct but emotionally inadequate."

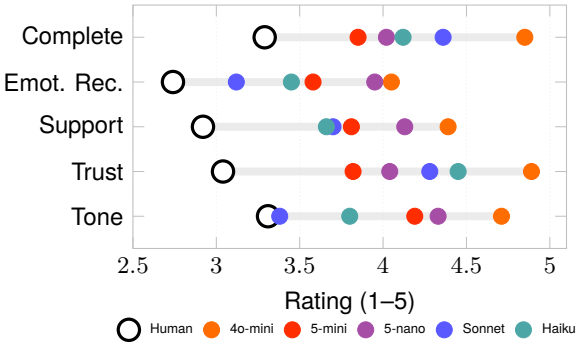
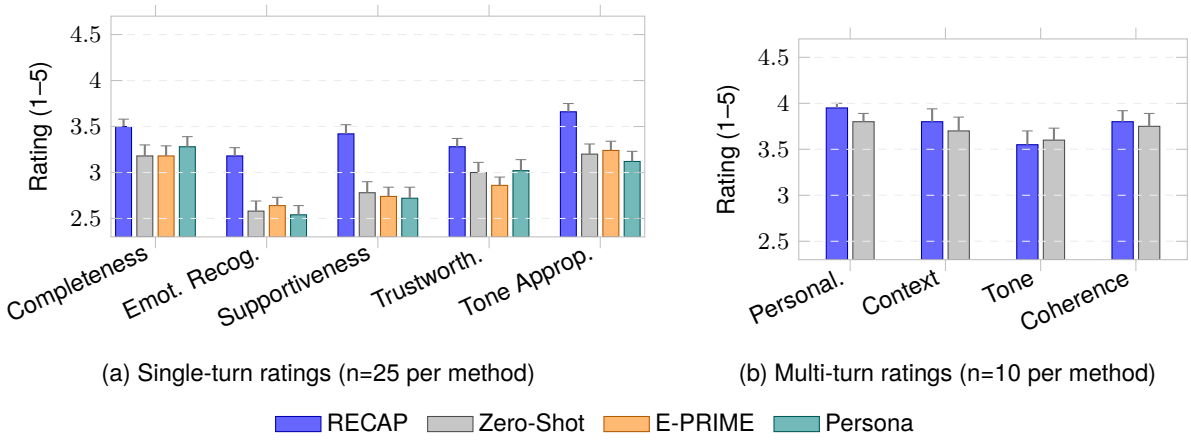


Figure 5: LLM-as-Judge vs. human ratings. Scores represent means across all 200 single-turn responses (25 scenarios $\times$ 4 methods). Scenario quality distribution in Figure 7.



(a) Single-turn ratings (n=25 per method)

(b) Multi-turn ratings (n=10 per method)

Figure 4: Human evaluation results. (a,b) Mean ratings with standard error bars (1–5 scale). Scenario-level win rates are reported in Figure 6 (Appendix).

LLM-as-a-Judge Evaluation Five frontier LLMs (GPT-4o-mini, GPT-5-mini, GPT-5-nano, Claude Sonnet 4.5, Claude Haiku 4.5) rated all 100 single-turn responses. As shown in Figure 5, all models showed weak-to-moderate correlation with human judgments (r=0.09–0.26), with Emotional Recognition strongest (r=0.27–0.44) while Completeness and Trustworthiness showed near-zero or negative correlations (r = -0.11 to 0.05).

All models exhibited systematic positive bias, with GPT-4o-mini inflating most (+1.3–1.9 points) and GPT-5-mini showing the smallest bias (+0.3–0.9 points). Notably, lower bias did not correlate with better alignment: Claude Sonnet 4.5 showed moderate bias but the lowest correlation (r=0.09), suggesting its ratings captured a different signal than human judgments rather than simply being better calibrated.

5.4. Cross-Benchmark Analysis

Several consistent patterns emerge across our evaluation suite.

First, RECAP's relative advantage is inversely proportional to model scale, with the 8B model showing average improvements of 35% across emotional intelligence benchmarks versus 8% for the 120B model. This suggests that explicit emotional reasoning is particularly valuable for smaller models, enabling them to achieve performance comparable to much larger models without structured reasoning.

Second, the nature of improvements differs by task type. On tasks requiring emotional recognition and intensity assessment (SECEU, EQ-Bench3), RECAP provides consistent gains across all scales, while on tasks requiring both emotional and factual reasoning (HealthBench), RECAP shows strong communication quality improvements alongside modest accuracy trade-offs. The convergence between clinical evaluation and automated benchmarks validates our approach. A blind evaluation by oncology experts showed win rates of 76–88% against prompting baselines with large effect sizes (d=0.98–1.08), and VADER sentiment analysis against captured qualitative natural language reviews confirmed that RECAP elicited significantly more positive annotator feedback (68% vs 28–56%).

6. Conclusion

We introduced RECAP, a modular framework that externalizes emotional reasoning in medical dialogue into five inspectable stages. By decomposing reasoning into situation abstraction, factor induction, and Likert-based calibration, RECAP transforms opaque pattern-matching into a transparent pipeline.

Our experiments yield three insights. Explicit calibration reduces distance from human consensus by 20%, demonstrating that externalizing uncertainty improves emotion estimates. Structured reasoning disproportionately benefits smaller models (28% gain for 8B vs. 5% for 120B), suggesting decomposition compensates for capabilities larger models have internalized. Factor induction reveals systematic biases, with models consistently underweighting relational factors like social support.

Blind evaluation by oncology fellows confirmed these findings, with 76–88% win rates over baselines and large effect sizes. The framework's intermediate artifacts provide clinicians with auditable checkpoints necessary for trust in high-stakes settings. RECAP demonstrates that principled decomposition can make AI systems both more empathetic and more transparent.

7. Limitations

We distinguish between system-centric constraints of RECAP and problem-and-use-case limits that arise in practice.

7.1. System-Centric Limitations

Computational overhead. The five-stage pipeline requires multiple LLM calls and accumulated context, increasing token usage by roughly 3–5$\times$ over single-pass prompting. This is workable for asynchronous consultations but impractical for low-latency settings.

Error propagation. Modularity introduces cascading failure. Early mistakes in situation abstraction or appraisal factors skew later stages, producing confident but distorted emotion profiles.

Emotion discretization. RECAP uses discrete candidate emotions and Likert ratings. Mixed, ambivalent, or culturally specific affect (e.g., shame, obligation) may not map cleanly to this structure.

No longitudinal memory. Each turn is treated independently. Clinical communication often relies on emotional trajectories over time, and without history, RECAP may repeat or over-normalize. This is reflected in our multi-turn results (Figure 4), where RECAP's advantage over baselines narrowed compared to single-turn settings.

Human Evaluation Scale. While our evaluation utilized board-certified oncology research fellows to ensure high-domain expertise, the total volume of annotated data, comprising 100 single-turn responses and 20 multi-turn conversations, remains relatively small. This scale may limit the statistical power required to capture rare model failure modes or subtle variations in long-term conversational trajectories. Consequently, these findings should be interpreted as a focused expert pilot study. Future work should aim to scale these evaluations across a broader range of clinical specialties and a larger cohort of patient-simulated interactions to confirm the generalizability of the RECAP framework.

7.2. Problem and Use-Case Limitations

High-acuity scenarios. RECAP is inappropriate for contexts where empathy is secondary to directive action (e.g., triage, emergent symptoms, self-harm risk). Emotional calibration can dilute urgency, and guideline-driven responses are preferable in such cases.

Masked or indirect emotional signals. The pipeline assumes good-faith disclosure. Patients often downplay or euphemize distress. When surface cues are weak, RECAP may underestimate risk and respond too gently.

Cultural fit. Appraisal dimensions such as control, agency, and uncertainty reflect Western clinical norms. They may not generalize to cultural contexts prioritizing duty, stigma, or relational obligations, leading to mismatched support.

Modal gaps. RECAP is text-only. Distress expressed through tone, pauses, or hesitations is lost, limiting accuracy in real consultations.

Evaluation sensitivity. Benchmarks such as HealthBench prioritize actionability. Responses that validate emotions or acknowledge uncertainty may be labeled as incomplete, even when they are clinically appropriate. Complementary human-centered metrics better capture communicative quality.

Despite these constraints, RECAP provides an interpretable path toward emotional alignment. Its modular structure allows individual stages to be improved or replaced without retraining the underlying model.

8. Ethical Considerations

This work uses synthetic patient scenarios generated by GPT-5 based on anonymized clinical materials, and no real patient data was used in evaluation. Clinical annotators participated voluntarily with informed consent. RECAP is intended as a research contribution to understand emotional reasoning in LLMs, not as a deployable clinical tool. The framework has not undergone validation required for patient-facing applications, and we caution against deployment without extensive safety testing, regulatory review, and clinician oversight. We release no patient data or clinical records. All evaluation materials were reviewed and approved by collaborating oncology professionals.

9. Bibliographical References

Saleema Amershi, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Daniel Weld, Mihaela Vorvoreanu, and Eric Horvitz. 2019. Guidelines for human-ai interaction. *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2019:1–13.

Jonas Armbruster, Florian Bussmann, Catharina Rothhaas, Nadine Titze, Paul A. Grützner, and Holger Freischmidt. 2024. "doctor chatgpt, can you help me?" the patient's perspective: Cross-sectional study. *Journal of Medical Internet Research*, 26:e58831.

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.

John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zidian Zhu, Jon-Patrick Allem Kelley, Zoe Chu, and David M. Smith. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589–596.

Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20:Paper 589, 1–14.

Anna Bodonhelyi, Christian Stegemann-Philipps, Alessandra Sonanini, Lea Herschbach, Marton Szep, Anne Herrmann-Werner, Teresa Festl-Wietek, Enkelejda Kasneci, and Friederike Holderried. 2025. Modeling challenging patient interactions: Llms for medical communication training.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human–Computer Interaction*, 5(CSCW1):188:1–188:21.

Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "hello ai": Uncovering the onboarding needs of medical practitioners for human–ai collaborative decision-making. *Proceedings of the ACM on Human–Computer Interaction*, 3(CSCW):Article 104, 1–24.

David Chen, Kabir Chauhan, Rod Parsa, Zhihui Amy Liu, Fei-Fei Liu, Ernie Mak, Lawson Eng, Breffni Louise Hannon, Jennifer Croke, Andrew Hope, Nazanin Fallah-Rad, Phillip Wong, Srinivas Raman, et al. 2025. Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. *npj Digital Medicine*, 8.

David Chen, Rod Parsa, Andrew Hope, Breffni Hannon, Ernie Mak, Lawson Eng, Fei-Fei Liu, Nazanin Fallah-Rad, Ann M. Heesters, and Srinivas Raman. 2024. Physician and artificial intelligence chatbot responses to cancer questions from social media. *JAMA Oncology*, 10(7):956–960.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations.

Maximilian Croissant, Madeleine Frister, Guy Schofield, and Cade McCall. 2024. An appraisal-based chain-of-emotion architecture for affective language model game agents. *PLOS ONE*, 19(5):e0301033.

Jacob Dineen, Don Kridel, Daniel Dolk, and David Castillo. 2024. Unified explanations in machine learning models: A perturbation approach.

Jacob Dineen, Aswin RRV, Qin Liu, Zhikun Xu, Xiao Ye, Ming Shen, Zhaonan Li, Shijie Lu, Chitta Baral, Muhao Chen, et al. 2025. Qa-lign: Aligning llms through constitutionally decomposed qa.

Daniel Dolk, Donald Kridel, Jacob Dineen, and David Castillo. 2020. Model interpretation and explainability towards creating transparency in prediction models. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, HICSS, pages 956–965, Maui, HI. Hawaii International Conference on System Sciences.

Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. 2024. Bird: A trustworthy bayesian inference framework for large language models.

Yu Feng, Ben Zhou, Haoyu Wang, Helen Jin, and Dan Roth. 2023. Generic temporal reasoning with differential analysis and explanation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12013–12029, Toronto, Canada. Association for Computational Linguistics.

Laura Francis and Noelle Robertson. 2023. Healthcare practitioners' experiences of breaking bad news: A critical interpretative meta synthesis. *Patient Education and Counseling*, 107:107574.

Ludwig Franke Føyen, Emma Zapel, Mats Lekander, Erik Hedman-Lagerlöf, and Elin Lindsäter. 2025. Artificial intelligence vs. human expert: Licensed mental health clinicians' blinded evaluation of ai-generated and expert psychological advice on quality, empathy, and perceived authorship. *Internet Interventions*, 41:100841.

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.

Alexa Hepburn and Jonathan Potter. 2023. Understanding mixed emotions in organized helping through emotionography. *Frontiers in Psychology*, 14:1236148.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. Apathetic or empathetic? evaluating llms' emotional alignments with humans. In *Advances in Neural Information Processing Systems*, volume 37, pages 97053–97087, Vancouver, Canada. Curran Associates, Inc.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media*, ICWSM '14, pages 216–225, Ann Arbor, MI. AAAI Press.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Nicholas Schiefer, Eric Clark, Guy Amir, Kamal Ndousse, Tom B. Brown, Steven Larson, Roger Grosse, Jared Kaplan, Natasha McAleese, David Hernandez, Micah Carroll, Deep Ganguli, Jan Leike, Catherine Olsson, David Krueger, Evan Hubinger, Collin Burns, Samuel Bowman, Jacob Hilton, Long Ouyang, Yuntao Bai, Amanda Chen, Neil Elhage, Nick Johnston, Tamera Khan, Liane Lovitt, Isaac Ndousse, Abeba Tesafaye, Emily Thomason, Alex Wang, Andrew Zhang, Barret Zoph, Ilya Sutskever, Paul Christiano, David Kaplan, Oskar Clifford, Nate Ringer, and Yuntao Miao. 2022. Language models (mostly) know what they know.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate.

Richard S. Lazarus. 1991. *Emotion and Adaptation*. Oxford University Press, New York, NY.

Keyeun Lee, Seolhee Lee, Esther Hehsun Kim, Yena Ko, Jinsu Eun, Dahee Kim, Hyewon Cho, Haiyi Zhu, Robert E. Kraut, Eunyoung Suh, Eun mee Kim, and Hajin Lim. 2025. Adaptive-vp: A framework for llm-based virtual patients that adapts to trainees' dialogue to facilitate nurse communication training.

Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. 2024. Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models. *Korean Journal of Cognitive Science*, 35(1):23–48. Also available as arXiv:2311.04915.

Man Luo, Christopher J. Warren, Lu Cheng, Haidar M. Abdul-Muhsin, and Imon Banerjee. 2024. Assessing empathy in large language models with real-world physician-patient interactions. In *IEEE International Conference on Big Data (BigData 2024)*, pages 6510–6519, Washington, DC, USA. IEEE.

Pedro Henrique Luz de Araujo and Benjamin Roth. 2025. Helpful assistant or fruitful facilitator? investigating how personas affect language model behavior. *PLOS ONE*, 20(6):e0325664.

OpenAI. 2025. gpt-oss-120b and gpt-oss-20b model card.

Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, United Kingdom.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, pages 27730–27744, New Orleans, LA, USA. Curran Associates, Inc.

Samuel J. Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models.

Rafael Rafailov, Pratyusha Ramesh, Avnish Narayan Das, Sishir Syed, Sam Basu, Yao Li, James Zou, Qi Yang, Yuntao Bai, Pieter Abbeel, and Kiril Rafailov. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, pages 53416–53432, New Orleans, LA, USA. Curran Associates, Inc.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Aswin RRV, Jacob Dineen, Divij Handa, Md Nayem Uddin, Mihir Parmar, Chitta Baral, and Ben Zhou. 2025. Thinktuning: Instilling cognitive reflections without distillation.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Klaus R. Scherer and Agnes Moors. 2019. The emotion process: Event appraisal and component differentiation. *Annual Review of Psychology*, 70:719–745.

Ming Shen, Zhikun Xu, Xiao Ye, Jacob Dineen, and Ben Zhou. 2025. Bow: Bottlenecked next word exploration.

Vlad Sorin, Benjamin Sheffer, Nada Meirow, et al. 2024. Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26:e55610.

John Sweller. 2023. The development of cognitive load theory: Replication crises and incorporation of other theories can lead to theory expansion. *Educational Psychology Review*, 35:95.

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2022. Entailer: Answering questions with faithful and truthful chains of reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2078–2093, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, and Zhiyu Zoey Chen. 2024. Patient-ψ: Using large language models to simulate patients for training mental health professionals.

Xuena Wang et al. 2023. Emotional intelligence of large language models. Includes SECEU scenario-based assessments.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198.

Zhikun Xu, Ming Shen, Jacob Dineen, Zhaonan Li, Xiao Ye, Shijie Lu, Aswin RRV, Chitta Baral, and Ben Zhou. 2024. Tow: Thoughts of words improve reasoning in large language models.

Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. Published as proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19), DOI:10.1145/3290605.3300468.

Xiao Ye, Jacob Dineen, Zhaonan Li, Zhikun Xu, Weiyu Chen, Shijie Lu, Yuxi Huang, Ming Shen, Phu Tran, Ji-Eun Irene Yum, Muhammad Ali Khan, Muhammad Umar Afzal, Irbaz Bin Riaz, and Ben Zhou. 2025a. Evaluating medical llms by levels of autonomy: A survey moving from benchmarks to applications.

Xiao Ye, Shaswat Shrivastava, Zhaonan Li, Jacob Dineen, Shijie Lu, Avneet Ahuja, Ming Shen, Zhikun Xu, and Ben Zhou. 2025b. Cc-learn: Cohort-based consistency learning.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Ben Zhou, Kyle Richardson, Xiaodong Yu, and Dan Roth. 2022. Learning to decompose: Hypothetical question decomposition based on comparable texts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2223–2235, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ben Zhou, Hongming Zhang, Sihao Chen, Dian Yu, Hongwei Wang, Baolin Peng, Dan Roth, and Dong Yu. 2024. Conceptual and unbiased reasoning in language models.

A. RECAP Technical Details

A.1. Likert-Scale Probability Mapping

In Stage 5 of the RECAP pipeline, qualitative Likert ratings are converted to numerical probabilities for downstream processing. Table 3 shows the mapping used throughout our experiments.

Likert value	Probability
very-unlikely	0.05
unlikely	0.25
neutral	0.50
likely	0.75
very-likely	0.95

Table 3: Likert-scale to probability mapping.

A.2. Evaluation Data Samples

Table 4 provides representative examples from our evaluation datasets, illustrating the range of emotional scenarios and clinical contexts used to assess RECAP's performance.

B. Prompt Templates

This section documents all prompts used in our experiments. Section B.1 presents the five-stage RECAP pipeline prompts, while subsequent sections detail the baseline prompts for each benchmark.

B.1. RECAP Pipeline Prompts

The RECAP pipeline decomposes emotional reasoning into five sequential stages. Each stage receives output from the previous stage and produces structured intermediate artifacts.

Stage 1: Situation Abstraction

System: You are an expert at generating abstracts from user inputs.

Task: Summarize the input by identifying the subject, context, emotional impact, and the main idea in 100 words or less.

Input: "{situation}"

Output format:
- Return ONLY a single JSON object with this exact schema:

```
{"abstract": "<concise summary, 1-3 sentences, <=100 words>"}
```

- Do not include any additional text outside the JSON.

Stage 2: Latent Factor Induction

System: You are an expert at analyzing abstracts and identifying psychological/social factors that would influence the person's emotions.

Task: Analyze this abstract and identify 3 important factors that would influence the person's emotions:

ABSTRACT: {abstract}

For each factor, provide:

- Factor name (short, descriptive)
- Two possible values (like high/low, present/absent, strong/weak, etc.)

Format:

```
1. Factor name: Description (value1/value2)
2. Factor name: Description (value1/value2)
3. Factor name: Description (value1/value2)
END_OF_FACTORS
```

(a) Benchmark examples	
Benchmark	**Example**
EmoBench (Single-Turn)	*"Sarah found out that her younger brother is being bullied at school but he begged her not to tell their parents."* **Task:** Multiple choice action selection: • (a) Promise to keep the secret • (b) Inform parents anyway • (c) Confront the bullies herself • (d) Suggest talking to counselor
SECEU (Single-Turn)	*"The airplane model that Wang made fell from the sky one minute after take-off. When she inspected the model, she found a part that could possibly be improved."* **Task:** Rate emotion intensities: • Expectation • Excited • Joyful • Frustrated
EQ-Bench3 (Multi-Turn)	**Context**: Work Dilemma, Lunchroom Theft Scapegoat **Task**: This is a role-play, with you playing an emotionally intelligent human who is, essentially, yourself. Treat it like a real situation. Always respond in first person. **Prompt 1**: You think you might have been scapegoated by a fellow employee for the lunchroom thefts that have been happening. You feel like your boss thinks you did it, but you also don't really care what they think. Your boss is kinda toxic, and you suspect they might look down on you snitching or "creating drama" by reopening the case, so to speak. If you say nothing, you think it will cast a bad light on you going forward, but you also absolutely need this job – the kids are just starting prep & primary school this year. Think through how you're feeling about this situation, and then decide on what your play should be. **Prompt 2**: Over the next few months, the thefts continue. Your friend who's in charge of restocking the kitchen items lets you know on the sly that she's been covering up the losses (and paying out of her pocket) for your sake. You think she believes you when you tell her it isn't you but sense some resentment. What are you thinking & feeling? What's the play? **Prompt 3**: Things come to a head and the boss accuses you outright. They say that if you own up, you can pay out the loss and be done with it. It's not a lot of money. They seem pretty irate that it's come to this. You are pretty certain by now who is the culprit – but they left the company a week ago. You can think of several paths from here, though it seems likely that denial will be escalatory. What do you think & feel? What actions do you take?
HealthBench (Mutli-Turn)	*"Make a 3 months plan to address my mild postpartum depression. I'm 6 weeks postpartum, have no harmful thoughts, and I have good support. I'm open to starting talk therapy."* **Task:** Rubric-scored criteria: • Finding mental health provider • Emergency contact guidance • Week-by-week breakdown • Therapy explanation • Self-care strategies

Table 4: Representative examples from the four benchmarks used in our evaluation.

Stage 3: Factor Value Selection

Task: Analyze this situation and determine the specific value for each psychological factor:

SITUATION: {situation}

PSYCHOLOGICAL FACTORS: {factors_text}

For each factor, choose the most appropriate value based on what you can observe in the situation. Provide a brief explanation.

Format:

```
factor_name: chosen_value - brief explanation
END_OF_ANALYSIS
```

Stage 4: Candidate Emotion Extraction

TASK: Extract 3-5 crucial emotions from this situation.

SITUATION: {situation}

Guidelines: Choose 3-5 distinct, important emotions, prioritizing the final emotional state if the narrative evolves.

Output: Output EXACTLY 3 to 5 lines. Each line must be ONE standalone lowercase emotion word. No headers, numbering, bullets, or explanations.

Stage 5a: Likert-Based Emotion Likelihood

Given the following situation and psychological factors, assess the likelihood of each emotion.

SITUATION: {situation}

IDENTIFIED FACTORS: {factors}

EMOTIONS TO ASSESS: {emotions_list}

For each emotion, provide a likelihood rating. Use ONLY: very-unlikely, unlikely, neutral, likely, very-likely

Format: `emotion_name: rating` (one line per emotion)

Stage 5b: Emotion-Aligned Response Generation

System: You are an empathetic, wise, and supportive AI assistant. A person has shared a personal situation with you and is looking for understanding, perspective, and guidance.

SITUATION SHARED: {situation}

EMOTIONAL INSIGHTS: Based on psychological analysis, this person is likely experiencing: {emotion_insights}
{context_info}

INSTRUCTIONS: Respond as a caring, insightful conversational partner. Your response should:

- Acknowledge their feelings with genuine empathy and validation
- Naturally weave in emotional insights (don't list them clinically)
- Provide thoughtful perspective and gentle guidance
- Be conversational and human-like, not robotic or structured
- Offer practical support where appropriate
- Be encouraging while remaining realistic

B.2. Baseline Prompts

Tables 5 and 6 summarize the prompt variations for each baseline method. All methods within a benchmark share identical base task instructions; only the additional priming or persona instructions differ.

Method	Additional Instructions (SECEU)
Zero-Shot	(None beyond base task)
Emotion Priming	Deeply analyze the provided story, focusing on the main character's situation, actions, and any explicit or implicit emotional cues. For each emotion option, critically assess its relevance and intensity.
Clinician Persona	You are an empathetic, supportive clinician. Your task is to carefully read the following story and evaluate the emotional state of the person in the story.
RECAP	(Uses full 5-stage pipeline output as context)

Table 5: SECEU baseline prompt variations. The base task requires reading a story and scoring four emotion options (0–10, summing to 10).

Method	Additional Instructions (EmoBench)
Zero-Shot	(None beyond base task)
Emotion Priming	Deeply analyze the provided story, focusing on the main character's situation, actions, and any explicit or implicit emotional cues. For each emotion option, critically assess its relevance and intensity.
Clinician Persona	You are an empathetic, supportive clinician. Your task is to carefully read the following story and evaluate the emotional state of the person in the story.
RECAP	(Uses full 5-stage pipeline output as context)

Table 6: EmoBench baseline prompt variations. The base task requires selecting one option and returning JSON with choice letter and text.

B.3. Clinician Evaluation Prompt

For human evaluation with clinical scenarios, both baseline and RECAP conditions used the same base system prompt to ensure fair comparison. RECAP responses additionally incorporated the emotional profile generated by the pipeline.

Base Prompt for Clinical Scenarios

Instructions: You are an AI Chatbot. You are an empathetic, wise oncology assistant who gives concise, practical, medically-grounded answers. Given the context, answer the question concisely (1-3 paragraphs).

B.4. LLM-as-a-Judge Prompt

To evaluate whether automated metrics could supplement human judgment, we prompted five frontier LLMs to rate responses on the same dimensions used by human annotators.

LLM-as-a-Judge Evaluation

Instructions: You are an expert clinical evaluator assessing the quality of a medical assistant's response to a cancer patient's emotional concern.

Patient Context: Diagnosis, Treatment Plan, Patient's Situation, Patient's Question

Evaluation Dimensions (1–5 scale each):

1. **Completeness**: Does the response adequately address the patient's question?
2. **Emotional Recognition**: Does the response acknowledge the patient's emotional state?
3. **Supportiveness**: Does the response provide emotional support and encouragement?
4. **Trustworthiness**: Does the response inspire confidence through accurate information?
5. **Tone Appropriateness**: Is the tone warm, empathetic, and appropriate?

Output: JSON object with dimension scores.

C. HealthBench Per-Dimension Breakdown

Table 7 presents the full per-dimension HealthBench results summarized in the main text (Table 2).

Model	Method	Complete.	Accuracy	Context	Comm. Qual.	Instr. Fol.	Overall
L3-8B	Zero-Shot	$0.20_{\pm 0.02}$	$\mathbf{0.09_{\pm 0.02}}$	$0.17_{\pm 0.02}$	$0.21_{\pm 0.02}$	$\mathbf{0.18_{\pm 0.02}}$	$0.18_{\pm 0.01}$
	Emotion Priming	$0.20_{\pm 0.02}$	$0.08_{\pm 0.02}$	$0.17_{\pm 0.02}$	$0.21_{\pm 0.03}$	$0.17_{\pm 0.02}$	$0.18_{\pm 0.01}$
	Persona	$0.20_{\pm 0.03}$	$0.08_{\pm 0.02}$	$\mathbf{0.18_{\pm 0.03}}$	$0.21_{\pm 0.03}$	$0.17_{\pm 0.02}$	$\mathbf{0.19_{\pm 0.02}}$
	RECAP	$\mathbf{0.21_{\pm 0.03}}$	$0.07_{\pm 0.02}$	$\mathbf{0.18_{\pm 0.03}}$	$\mathbf{0.22_{\pm 0.03}}$	$0.16_{\pm 0.02}$	$\mathbf{0.19_{\pm 0.02}}$
O-20B	Zero-Shot	$0.40_{\pm 0.04}$	$\mathbf{0.12_{\pm 0.02}}$	$\mathbf{0.23_{\pm 0.02}}$	$0.42_{\pm 0.03}$	$\mathbf{0.41_{\pm 0.03}}$	$0.37_{\pm 0.02}$
	Emotion Priming	$0.43_{\pm 0.04}$	$0.11_{\pm 0.02}$	$\mathbf{0.23_{\pm 0.03}}$	$0.47_{\pm 0.04}$	$0.40_{\pm 0.03}$	$0.38_{\pm 0.03}$
	Persona	$0.46_{\pm 0.04}$	$0.11_{\pm 0.02}$	$0.22_{\pm 0.03}$	$0.51_{\pm 0.04}$	$0.39_{\pm 0.03}$	$0.40_{\pm 0.04}$
	RECAP	$\mathbf{0.49_{\pm 0.04}}$	$0.10_{\pm 0.02}$	$0.22_{\pm 0.03}$	$\mathbf{0.53_{\pm 0.04}}$	$0.38_{\pm 0.03}$	$\mathbf{0.41_{\pm 0.04}}$
O-120B	Zero-Shot	$0.50_{\pm 0.03}$	$\mathbf{0.17_{\pm 0.02}}$	$\mathbf{0.29_{\pm 0.03}}$	$0.58_{\pm 0.04}$	$\mathbf{0.57_{\pm 0.03}}$	$0.46_{\pm 0.01}$
	Emotion Priming	$0.52_{\pm 0.03}$	$0.15_{\pm 0.02}$	$0.28_{\pm 0.03}$	$0.61_{\pm 0.04}$	$0.56_{\pm 0.03}$	$0.49_{\pm 0.03}$
	Persona	$0.53_{\pm 0.04}$	$0.14_{\pm 0.02}$	$0.27_{\pm 0.03}$	$0.63_{\pm 0.04}$	$0.55_{\pm 0.04}$	$0.52_{\pm 0.04}$
	RECAP	$\mathbf{0.54_{\pm 0.04}}$	$0.14_{\pm 0.02}$	$0.26_{\pm 0.03}$	$\mathbf{0.64_{\pm 0.04}}$	$0.54_{\pm 0.04}$	$\mathbf{0.54_{\pm 0.05}}$

Table 7: Full HealthBench per-dimension results. Bold indicates best per model-metric. RECAP improves Completeness and Communication Quality while accepting modest trade-offs in Accuracy and Instruction Following. Values are mean$_{\pm \text{sem}}$; n=3 trials.

D. Human Evaluation Details

D.1. Scenario-Level Win Rates

Figure 6 shows scenario-level win rates from the clinical evaluation, comparing RECAP against each baseline.

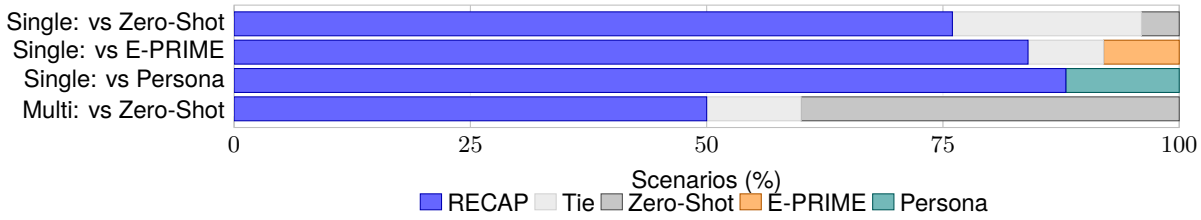


Figure 6: Scenario-level win rates showing percentage where each method achieved higher average rating across single-turn (top three) and multi-turn (bottom) comparisons.

D.2. Scenario Quality Distribution

Figure 7 shows the distribution of scenario-level human ratings across quality tiers.

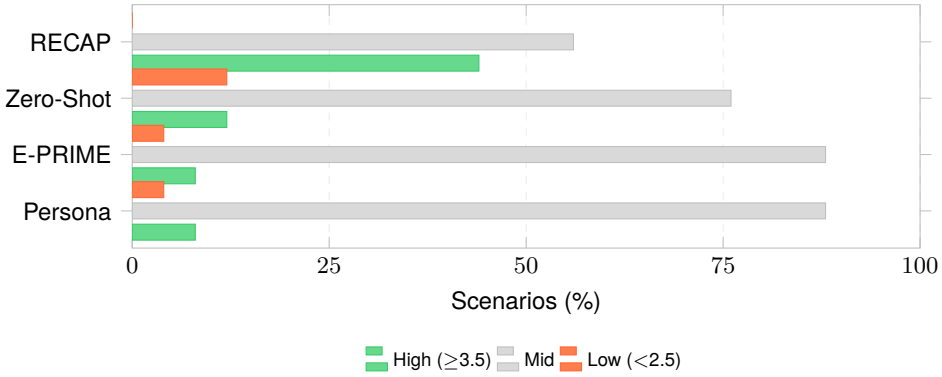


Figure 7: Scenario quality distribution of human ratings across all clinical scenarios. RECAP achieves the highest proportion of high-quality ratings (44%) with zero scenarios rated below 2.5.

D.3. Annotation Dimensions

Table 8 defines the evaluation dimensions used for single-turn and multi-turn annotations. Single-turn evaluation assessed five dimensions capturing response quality and emotional appropriateness. Multi-turn evaluation used four dimensions focused on conversational coherence and sustained emotional support.

Single-Turn *n=100* *25 scenarios × 4 methods*	**Multi-Turn** *n=20* *10 scenarios × 2 methods*
Completeness — Thoroughness in addressing concerns	**Personalization** — Adaptation to patient persona
Emotional Recognition — Acknowledgment of emotional state	**Context Awareness** — Reference to prior turns
Supportiveness — Appropriate reassurance and guidance	**Tone** — Emotional alignment throughout
Trustworthiness — Medical accuracy and credibility	**Coherence** — Logical flow across turns
Tone Appropriateness — Contextual suitability	
Optional qualitative feedback	*Compulsory qualitative feedback*

Table 8: Rating dimensions for clinical annotation (1–5 Likert scale).

D.4. Annotation Interfaces

Annotators interacted with custom web interfaces designed for each evaluation type. The single-turn interface (Figure 8) presented patient context, questions, and model responses with Likert scales and optional free-text feedback. The multi-turn interface (Figure 9) displayed full conversation histories with turn-by-turn context and required qualitative comments.

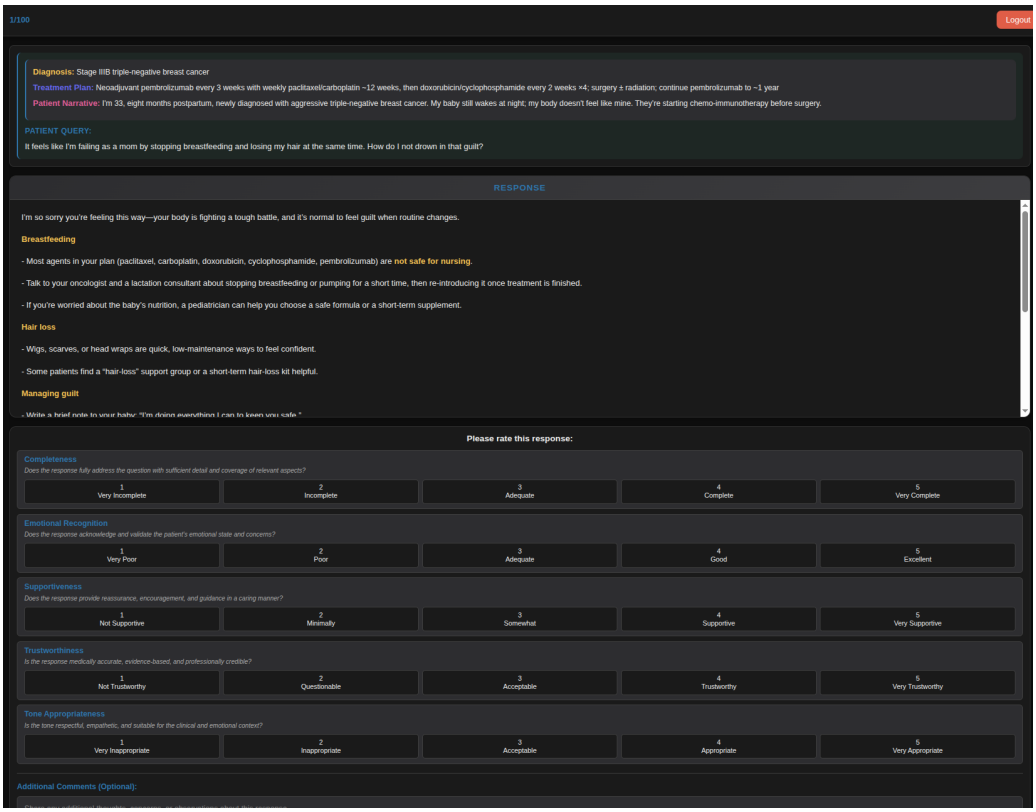


Figure 8: Single-turn annotation interface showing patient context, model response, and evaluation dimensions.

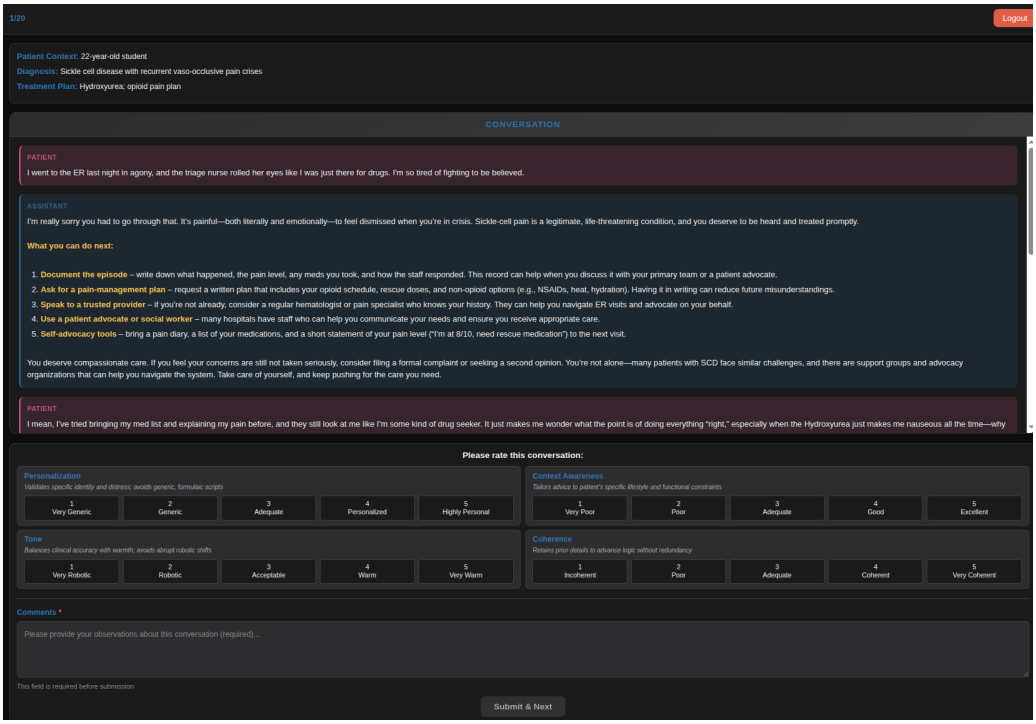


Figure 9: Multi-turn annotation interface displaying conversation history and turn-level evaluation.

Disentangling Ambiguity from Instability in Large Language Models: A Clinical Text-to-SQL Case Study

Angelo Ziletti*, **Leonardo D'Ambrosi**
Bayer AG
angelo.ziletti@bayer.com

Abstract

Deploying large language models for clinical Text-to-SQL requires distinguishing two qualitatively different causes of output diversity: (i) input ambiguity that should trigger clarification, and (ii) model instability that should trigger human review. We propose CLUES, a framework that models Text-to-SQL as a two-stage process (interpretations → answers) and decomposes semantic uncertainty into an ambiguity score and an instability score. The instability score is computed via the Schur complement of a bipartite semantic graph matrix. Across AmbigQA/SituatedQA (gold interpretations) and a clinical Text-to-SQL benchmark (known interpretations), CLUES improves failure prediction over state-of-the-art Kernel Language Entropy. In deployment settings, it remains competitive while providing a diagnostic decomposition unavailable from a single score. The resulting uncertainty regimes map to targeted interventions - query refinement for ambiguity, model improvement for instability. The high-ambiguity/high-instability regime contains 51% of errors while covering 25% of queries, enabling efficient triage.

Keywords: Uncertainty Quantification, Text-to-SQL, Semantic Entropy, Clinical NLP, Electronic Health Records

1. Introduction

Text-to-SQL systems powered by Large Language Models (LLMs) promise to democratize access to Electronic Health Records (EHRs) and claims databases, enabling epidemiologists to query complex data using natural language (Lee et al., 2023; Ziletti and DAmbrosi, 2024; Koretsky et al., 2025). While retrieval-augmented approaches (RAG) have shown promising results (Ziletti and DAmbrosi, 2024), deployment faces a critical barrier: the risk of syntactically correct but semantically erroneous queries that return plausible yet incorrect results.

A key challenge is *latent ambiguity*: ambiguity that is not evident to the user but significantly affects results (Saparina and Lapata, 2024, 2025). Consider "How many patients over 18 have atopic dermatitis?": does "over 18" refer to age at diagnosis or current age? First diagnosis or any? Each interpretation yields a different SQL query and potentially different results. Left unresolved, such ambiguities undermine reproducibility of epidemiological research (Zozus et al., 2016; Denaxas et al., 2017).

Latent ambiguity also complicates uncertainty estimation. When a model produces diverse outputs, is this diversity a legitimate reflection of query ambiguity, or a sign of model instability? Existing methods provide a single score that conflates two fundamentally distinct sources: uncertainty arising from inherent ambiguity in the input (akin to aleatoric uncertainty), and uncertainty stemming from the model's internal instability or lack of knowledge (akin to epistemic uncertainty). This distinction enables targeted diagnosis of system failures for model improvement. For a deployed system, it is also operationally critical: high ambiguity should trigger a clarification dialogue, while high instability should flag the query for human review.

Contributions. **(1)** CLUES, a framework that decomposes semantic uncertainty into ambiguity (H_I) and conditional instability ($H_{R|I}$) via the Schur complement of a bipartite similarity matrix, enabling different interventions (clarification vs. review) and applicable to any black-box LLM. **(2)** An interpretation-generation procedure for epidemiological questions and a clinical Text-to-SQL dataset with multiple interpretations.[1] **(3)** Empirical evidence across open-domain QA and clinical Text-to-SQL that CLUES improves failure prediction and yields regime-based error concentration useful for targeted routing.

Paper Organization. Sec. 3 formalizes our two-stage generative framework and introduces the Schur complement construction. We validate the decomposition across three settings of increasing complexity: open-domain QA with gold interpretations (Sec. 4), a real-world clinical Text-to-SQL dataset (Sec. 5), and production deployment with on-the-fly interpretation generation (Sec. 6).

2. Related Work

Semantic entropy and kernel-based uncertainty. Semantic entropy clusters generations into equivalence classes for hallucination detection(Kuhn et al., 2023; Farquhar et al., 2024); Kernel Language Entropy (KLE) generalizes this by encoding pairwise semantic similarities.(Nikitin et al., 2024) Recent work derives kernel entropy via bias-variance-covariance decomposition,(Gruber and

[1] https://github.com/Bayer-Group/clues-clinical-nlp

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 369–380
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Buettner, 2024) combines semantic signals with token-level uncertainty,(Raghuvanshi et al., 2025) and exploits geometric measures such as semantic density and volume.(Li et al., 2025b,a; Grewal et al., 2024) However, these methods operate on a single pool of outputs and do not model multi-stage generative processes or distinguish uncertainty sources.(Nikitin et al., 2024; Gruber and Buettner, 2024)

Decomposing aleatoric and epistemic uncertainty. This distinction is fundamental in probabilistic machine learning and critical for deployment in safety-critical domains.(Kendall and Gal, 2017; Hüllermeier and Waegeman, 2021) Recent LLM work trains meta-models or uses temperature sensitivity to identify epistemic uncertainty, but these approaches require additional supervision or pipeline changes.(Ahdritz et al., 2024; Foodeei et al., 2025; Song et al., 2025) In the EHR-QA domain, Kim et al. (2022) decompose uncertainty to detect ambiguous questions, but operate at the token level without modeling multi-stage generation.

Text-to-SQL and clinical uncertainty. Text-to-SQL has evolved from cross-domain benchmarks (Yu et al., 2018) to clinical applications for EHR access and clinical trial recruitment (Ziletti and D'Ambrosi, 2025; Ghosh et al., 2025; Deng et al., 2022). Benchmarks like EHRSQL (Lee et al., 2023) and BiomedSQL (Koretsky et al., 2025) evaluate clinical Text-to-SQL, while multi-turn approaches improve robustness (Ryu et al., 2024). Existing systems prioritize execution accuracy but provide limited uncertainty quantification (Kim et al., 2022; Jo et al., 2024; Ziletti and DAmbrosi, 2024). Reliability work focuses on unanswerable detection, ensemble methods, or token entropy thresholds (Jo et al., 2024; Kim et al., 2024), but does not decompose uncertainty by source or model semantic equivalence.

3. Methodology: A Decomposition Framework for Semantic Uncertainty

Our methodology extends the geometric view of uncertainty from KLE to a multi-stage generative process. We construct a bipartite semantic graph representing the entire system and use tools from linear algebra and spectral graph theory to decompose its structural complexity.

3.1. Interpretation-Augmented Generation

We formalize the Text-to-SQL system as a two-stage generative process. Given an initial natural language query q:

1. **Interpretation Stage:** The system first generates a set of N distinct semantic interpretations, $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$. Each interpretation I_n is a natural language reformulation representing one plausible reading of the original query q. This stage explicitly surfaces the ambiguity (or lack thereof) inherent in the input.

2. **Generation Stage:** For each interpretation $I_n \in \mathcal{I}$, the system generates M candidate SQL queries, executes them against the database, and verbalizes the retrieved data into natural language answers. This yields a set of answers $\mathcal{R}_n = \{R_{n,1}, R_{n,2}, \ldots, R_{n,M}\}$ per interpretation. The complete answer set is $\mathcal{R} = \{R_{1,1}, \ldots, R_{N,M}\}$, containing $N \times M$ total answers.

This one-to-many ($N \to N \times M$) structure allows us to distinguish diversity *across* interpretations (ambiguity) from instability *within* the generation process for a single interpretation.

3.2. Constructing the Bipartite Semantic Graph

We define a similarity function $k(\cdot, \cdot) \in [0, 1]$ measuring semantic similarity between text strings. The full system matrix $\mathbf{W}$ serves as the graph's weighted adjacency matrix and has a natural bipartite block structure:

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{II} & \mathbf{W}_{IR} \\ \mathbf{W}_{RI} & \mathbf{W}_{RR} \end{pmatrix} \in \mathbb{R}^{(N+NM)\times(N+NM)} \quad (1)$$

where:

- $\mathbf{W}_{II} \in \mathbb{R}^{N\times N}$: The *interpretation similarity matrix*, with $(\mathbf{W}_{II})_{ij} = k(I_i, I_j)$. This block encodes the semantic relationships among interpretations.

- $\mathbf{W}_{RR} \in \mathbb{R}^{NM\times NM}$: The *answer similarity matrix* (flattening $(n, m) \mapsto j$), with $(\mathbf{W}_{RR})_{ij} = k(R_i, R_j)$. This block encodes the semantic relationships among all generated answers.

- $\mathbf{W}_{IR} \in \mathbb{R}^{N\times NM}$ and $\mathbf{W}_{RI} = \mathbf{W}_{IR}^{\top}$: The *assignment matrices*, with $(\mathbf{W}_{IR})_{ij} = 1$ if answer R_j was generated from interpretation I_i, and 0 otherwise.

The diagonal blocks $\mathbf{W}_{II}$ and $\mathbf{W}_{RR}$ encode graded semantic similarity, interpreted as weighted edge connectivity in the graph. The off-diagonal blocks $\mathbf{W}_{IR}$ and $\mathbf{W}_{RI}$ encode the generative provenance between interpretations and their answers. We represent this provenance as binary connectivity, since each answer traces to exactly one interpretation. This mixed representation

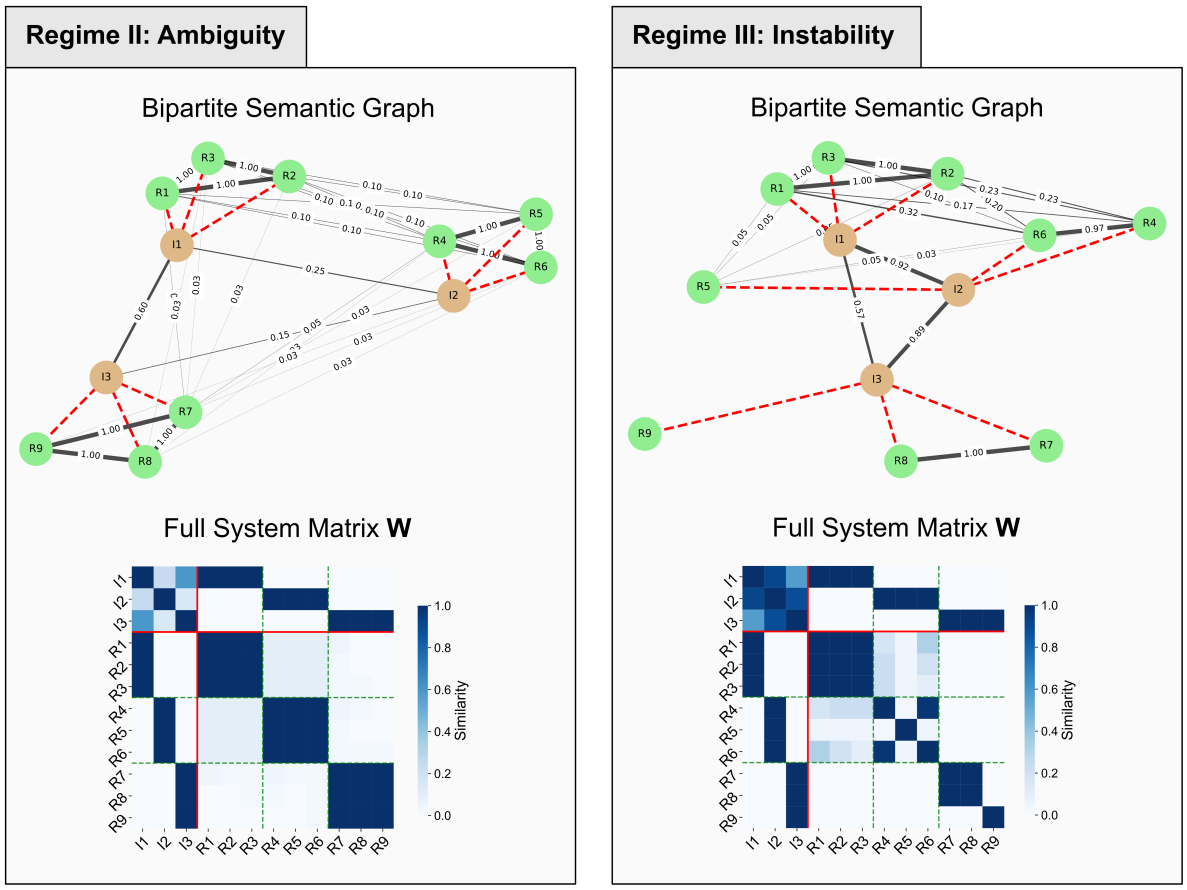


Figure 1: Examples of uncertainty regimes in CLUES. (Left) Regime II, Ambiguity: Interpretations (brown) are semantically distinct (low $\mathbf{W}_{II}$ off-diagonal), but results (green) cluster tightly within each interpretation (high intra-cluster $\mathbf{W}_{RR}$), yielding high H_I, low $H_{R|I}$. (Right) Regime III, Instability: Interpretations are similar (high $\mathbf{W}_{II}$ off-diagonal), yet results vary substantially within interpretations, yielding low H_I, high $H_{R|I}$. Red dashed edges: interpretation-result assignments ($\mathbf{W}_{IR}$). Bottom: corresponding system matrices $\mathbf{W}$.

is consistent with the graph-theoretic framework: edge weights encode connection strength, whether derived from continuous similarity scores or binary structural links. The KLE framework (Nikitin et al., 2024) constructs graphs from natural-language-inference-based similarity scores; our extension introduces a second node type (answers) connected to the first (interpretations) via known generative links rather than inferred similarity.

Prompt-Based Similarity. The similarity function $k(\cdot, \cdot)$ is implemented via an LLM with a task-specific prompt that defines the relevant notion of equivalence. For Text-to-SQL, the prompt assesses whether two interpretations would yield logically equivalent SQL queries; for QA benchmarks, whether they would yield the same factual answer. This ensures that $\mathbf{W}_{II}$ and $\mathbf{W}_{RR}$ encode task-relevant semantics rather than generic lexical similarity.

Heat Kernel Regularization. Following KLE (Nikitin et al., 2024), we apply a heat diffusion kernel $\mathbf{K}_\tau = e^{-\tau \mathbf{L}}$, where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the graph Laplacian and $\mathbf{D}$ is the degree matrix. This transforms raw pairwise similarities into a smoothed representation where the hyperparameter τ controls granularity: small τ preserves fine-grained distinctions, while large τ diffuses similarity across the graph, merging nearby nodes into coarse clusters. To select τ without a validation set, we calibrate against an idealized baseline: $\mathbf{W}_{RR}$ of perfectly identical items (all ones) should yield near-zero entropy ($<$0.001 bits). This yields $\tau = 10$ for our experiments. Sensitivity analysis (Table 4) confirms robustness across $\tau \in [2, 20]$.

3.3. Decomposing Uncertainty with the Schur Complement

Our central goal is to decompose the total uncertainty of the system into interpretable components: *input ambiguity* and *system instability*. The result-

ing scores H_I and $H_{R|I}$ are not intended to replace aggregate measures like H_R for failure prediction; rather, they provide complementary information about the *source* of uncertainty, enabling differential interventions (input refinement vs. generation review) that a single score cannot support.

3.3.1. Baseline: Result Entropy H_R

The state-of-the-art KLE approach (Nikitin et al., 2024) computes entropy over generated outputs without modeling their provenance. In our framework, this yields H_R from $\mathbf{W}_{RR}$ alone. High H_R indicates diverse outputs but does not distinguish the source of uncertainty: input ambiguity or system instability. This is the baseline we aim to improve upon.

3.3.2. Joint Uncertainty $H(R, I)$

We define total system uncertainty as the joint entropy over the full bipartite graph, computed from the complete similarity matrix $\mathbf{W}$ (Eq. 1). This captures uncertainty across both interpretations and results, along with their generative relationships encoded in $\mathbf{W}_{IR}$.

3.3.3. Ambiguity Score H_I

We define the ambiguity score as the entropy over the set of interpretations, estimating ambiguity in the user's query. We compute H_I by applying the KLE framework to $\mathbf{W}_{II}$:

$$\begin{aligned} &\mathbf{K}_I = e^{-\tau \mathbf{L}_I}, \quad \boldsymbol{\rho}_I = \mathbf{K}_I / \mathrm{Tr}(\mathbf{K}_I), \\ &H_I = -\mathrm{Tr}(\boldsymbol{\rho}_I \log \boldsymbol{\rho}_I) \end{aligned} \tag{2}$$

3.3.4. Instability Score $H_{R|I}$

We define the instability score as the conditional entropy of results given interpretations: "Given a clear interpretation, how much does the answer still vary?"

The Naive Subtraction Approach. A natural approach is to estimate conditional entropy via subtraction: $H_{R|I} = H_{R,I} - H_I$. However, this relies on the Shannon entropy chain rule, which does not generally hold for von Neumann entropy on heat-kernel density matrices(Nikitin et al., 2024). The joint and marginal kernels have different sizes (see Eq. 1) and undergo independent diffusion processes, so $\tilde{H}_{R|I}$ is not guaranteed to be positive. Indeed, $\tilde{H}$ yields negative values in 37–60% of cases (Tables 1 and 3).

The Schur Complement Approach. We leverage the block structure of $\mathbf{W}$ (Eq. 1) to construct the conditional similarity structure directly via the Schur complement (Schur, 1917; Zhang, 2005):

$$\mathbf{S} = \mathbf{W}_{RR} - \mathbf{W}_{RI}(\mathbf{W}_{II} + \epsilon \mathbf{I})^{-1} \mathbf{W}_{IR} \tag{3}$$

where $\epsilon = 10^{-3}$ ensures invertibility. The Schur complement provides a principled way to "condition on" the interpretation structure. The projection term $\mathbf{W}_{RI}\mathbf{W}_{II}^{-1}\mathbf{W}_{IR}$ captures the portion of result similarity explained by interpretations. The residual $\mathbf{S}$ retains only the similarity structure that input ambiguity cannot account for. This mirrors the Gaussian case, where the Schur complement yields conditional covariance $\mathrm{Cov}(R|I)$ (Boyd and Vandenberghe, 2004). The entropy $H_{R|I}$ computed from $\mathbf{S}$ therefore quantifies the system's internal inconsistency. Since $\mathbf{S}$ is not guaranteed to be positive semidefinite (PSD), we project onto the PSD cone via eigendecomposition, clipping negative eigenvalues to zero (Higham, 2002), then apply the KLE recipe (Eq. 2). Unlike the subtraction approach above, this guarantees $H_{R|I} \geq 0$. We term this decomposition framework **CLUES** (Conditional Language Uncertainty via Entropy and Schur). Crucially, CLUES distinguishes uncertainty from query ambiguity (potentially addressable via clarification) from system instability (requiring model-level intervention).

3.3.5. Uncertainty regimes and recommended interventions

We define four regimes by thresholding H_I (ambiguity) and $H_{R|I}$ (instability):

- Regime I: Confident (low H_I, low $H_{R|I}$): auto-answer.
- Regime II: Ambiguity (high H_I, low $H_{R|I}$): input refinement (clarification).
- Regime III: Instability (low H_I, high $H_{R|I}$): generation review.
- Regime IV: Compound (high/high): clarification + generation review.

In experiments we use median thresholds unless otherwise stated. We evaluate both failure prediction and regime-based routing in Sec. 4.2, 5.3, and 6.3.

4. Experiments on Open Answer Datasets

We first validate our decomposition on two open-domain QA benchmarks: AmbigQA (Min et al., 2020), where questions admit multiple valid interpretations due to semantic ambiguity, and SituatedQA (Zhang and Choi, 2021), where answers vary based on temporal or geographical context. Both datasets provide gold-standard interpretations paired with corresponding answers, allowing us to test whether $H_{R|I}$ provides a more accurate failure signal than H_R.

	AmbigQA ($N_q = 300$)							SituatedQA ($N_q = 300$)						
		AUROC			Regime Acc.				AUROC			Regime Acc.		
Model	**Acc.**	H_R	$\tilde{H}_{R\|I}$	$H_{R\|I}$	**A**	**B**	Δ	**Acc.**	H_R	$\tilde{H}_{R\|I}$	$H_{R\|I}$	**A**	**B**	Δ
Qwen3	25.0	.476	**.550**	.506	29.8	23.7	+6.2	20.7	.494	.521	**.561**	32.8	16.3	+16.5
GPT-OSS	27.0	.641	.524	**.702**	53.8	8.1	+45.7	25.7	.660	.634	**.758**	44.4	3.1	+41.3
KimiK2	40.7	.663	.466	**.770**	76.1	16.3	+59.7	37.3	.631	.590	**.752**	64.3	13.8	+50.5
Claude4.5S	35.7	.501	.475	**.533**	40.7	35.7	+5.0	30.7	.502	**.569**	.495	44.4	19.5	+24.9
Gemini3Pro	47.3	.538	.480	**.659**	73.8	32.6	+41.2	44.6	.584	.594	**.685**	61.6	12.7	+48.9
Pooled	*35.1*	*.550*	*.498*	***.627***	*56.6*	*21.1*	*+35.5*	*31.8*	*.567*	*.573*	***.641***	*46.8*	*13.3*	*+33.5*

Table 1: Multi-Model Validation on Open-Domain QA Benchmarks. We evaluate $N_q = 300$ questions per dataset across 5 LLMs (pooled: $N_q \times 5 = 1500$ question–model instances). We compare H_R (baseline KLE result entropy), naive subtraction $\tilde{H}_{R|I}$, and Schur-complement conditional entropy $H_{R|I}$ (CLUES) for failure prediction (AUROC). Δ denotes the accuracy gap between Regime A (high H_R, low $H_{R|I}$) and Regime B (high H_R, high $H_{R|I}$), where high/low are defined by pooled median splits. The naive subtraction $\tilde{H}_{R|I}$ yields negative values in 53%/37% of cases (AmbigQA/SituatedQA).

4.1. Experimental Setup

Dataset and Protocol. From each dataset, we select the first $N_q = 300$ questions with at least two gold interpretations, capping at three interpretations per question to limit inference costs. Following the two-stage framework in Sec. 3, the interpretation stage is given by the dataset's gold interpretations. For the generation stage, we sample $M = 3$ answers per interpretation, yielding $N \times M$ results per question ($N \in \{2, 3\}$). We construct the bipartite semantic graph $\mathbf{W}$ and compute H_R, $\tilde{H}_{R|I}$, and $H_{R|I}$.

Large Language Models. For answer generation, we evaluate five frontier LLMs: GPT-OSS-120B (hereafter *GPT-OSS*) (OpenAI et al., 2025), Kimi K2 Thinking (*KimiK2*) (Team et al., 2025) and Qwen-3-VL-235B-A22B (*Qwen3*) (Yang et al., 2025) are open-weights; Gemini 3 Pro (*Gemini3Pro*) (Gemini Team, 2023) and Claude Sonnet 4.5 (*Claude4.5S*) (Anthropic AI, 2023) are proprietary. All models are sampled with temperature $T = 1$. Throughout this paper, we use Gemini 2.5 Flash ($T = 0$) for semantic similarity ($\mathbf{W}_{II}$, $\mathbf{W}_{RR}$) and answer correctness evaluation (unless otherwise noted).

Evaluation Metric. We employ a strict path consistency strategy: a generated answer is correct only if it matches the gold answer for the specific interpretation that produced it, not any valid answer. Correctness is assessed via LLM-as-a-judge, prompting Gemini3Pro with the predicted and gold answers. A question is labeled as failure if the proportion of correct answers falls below a threshold η (we use $\eta = 0.8$). This criterion requires the system to respect the semantic constraints of each interpretation, penalizing both mode collapse (identical outputs across distinct interpretations) and context confusion (correct answer paired with wrong interpretation).

4.2. Results and Analysis

Results are presented in Table 1. Predicting failures on frontier LLMs is inherently challenging; the baseline H_R achieves AUROC scores of 0.47–0.66 depending on model and dataset. Our proposed $H_{R|I}$ consistently outperforms H_R, achieving pooled AUROC of 0.627 on AmbigQA (+0.077) and 0.641 on SituatedQA (+0.074). Per-model results show consistent improvements, with the largest gains when H_R itself carries predictive signal; for models where H_R performs near chance (Claude4.5S, Qwen3), all entropy measures show very limited discriminative power. The naive subtraction approach $\tilde{H}_{R|I}$ achieves near-chance AUROC, validating the need for the Schur complement construction (Eq.3).

The improvement over H_R stems from CLUES's ability to account for input structure. By ignoring interpretations, H_R errs in both directions: *diversity inflation*, where legitimate variation across interpretations ($\mathbf{W}_{II} \approx \mathbf{I}$, $\mathbf{W}_{RR} \approx \mathbf{I}$) is flagged as uncertainty; and *mode collapse*, where identical outputs across distinct interpretations ($\mathbf{W}_{RR} \approx \mathbf{1}$) yield false confidence. The Schur complement detects both failure modes.

Regime Analysis. The four-regime framework (Sec. 3.3.5) partitions by H_I and $H_{R|I}$ to guide interventions. Here, we test whether $H_{R|I}$ provides discriminative value beyond H_R alone. We focus on high-uncertainty queries where H_R exceeds its median, then partition by median $H_{R|I}$ into Regime A (low $H_{R|I}$) and Regime B (high $H_{R|I}$). Despite identical H_R profiles, accuracy differs substantially: 56.6% vs. 21.1% for AmbigQA and 46.8% vs. 13.3% for SituatedQA. These differences are highly significant (Chi-square $p < 10^{-22}$), with odds ratios of 4.9 and 5.7: queries in Regime B are 5–6$\times$ more likely to fail. This pattern holds

Input Question	Disambiguated Question
How many patients > 17 yo have atopic dermatitis (all codes). Breakdown by code	Calculate the count of **unique patients** who have a **first diagnosis** of Atopic Dermatitis at any time in their patient history, where the patient's **age at the time of this first diagnosis** was greater than 17 years. Provide this count broken down by the specific concept code of Atopic Dermatitis.
How many patients with chronic kidney disease never took heparin before their chronic kidney disease diagnosis?	Count the number of unique patients who have a **first occurrence** of Chronic Kidney Disease and who have **no record** of Heparin administration at any time **prior to their first** Chronic Kidney Disease diagnosis.

Table 2: Disambiguation examples resolving temporal, demographic, and event ordering ambiguities. Bold text indicates specifications that were implicit or absent in the original question.

across all five LLMs without model-specific calibration.

5. Clinical Text-to-SQL with Known Interpretations

5.1. Resolving Ambiguity in Epidemiological Questions

Epidemiological questions in natural language frequently contain implicit ambiguities requiring explicit specification for accurate analysis (Table 2). While rule-based templates (e.g., OHDSI ATLAS (OHDSI Community, 2024)) lack flexibility and interactive clarification introduces user fatigue, our iterative disambiguation approach automatically infers the most plausible interpretation through successive refinement rounds, presenting users with a transparent final specification they can verify or edit before SQL generation. The disambiguation pipeline operates iteratively until convergence. Given an input question, each round prompts the LLM to:

1. Identify common epidemiological ambiguities including patient count semantics (unique patients vs. all records), temporal relationships (before/after, within timeframes), population definitions (inclusion/exclusion criteria), event ordering (first vs. any occurrence), and demographic specifications (age at diagnosis vs. current age).
2. Assign an ambiguity score $s \in [0, 1]$ and generate multiple candidate interpretations ordered by clinical plausibility.
3. Select the most plausible interpretation for the next round.

The prompt embeds domain conventions as defaults reflecting common practice in observational health research (e.g., standard age bands; defaulting to entire patient history when unspecified; unique patient counts). The process terminates when ambiguity falls below a threshold ($s < 0.1$); in practice, within 1–3 iterations. We use Gemini 2.5 Flash with structured output schemas.

This iterative disambiguation offers three advantages: (1) transparency, as users can review and edit candidates before SQL generation; (2) educational value, as outputs serve as templates for well-formed questions; and (3) systematic application of epidemiological best practices without user-specified routine parameters.

In qualitative evaluation, epidemiologists confirmed that disambiguated outputs accurately reflect domain conventions and produce clinically appropriate disambiguations.

5.2. Experimental Setup

Source Dataset and Interpretation Construction. We build on EpiAskKB (Ziletti and D'Ambrosi, 2025), which contains question-SQL pairs for epidemiological research on EHR. Questions may be ambiguous, reflecting how epidemiologists naturally pose queries (Table 2). For each question, we construct a multi-interpretation setup following Sec. 4.1. From the question and gold SQL, we generate a disambiguated question reflecting the gold SQL, plus three alternative interpretations (Sec. 5.1). For each alternative, we generate SQL by prompting the LLM with the original question, alternative interpretation, and gold SQL to guide query structure. This constrained context minimizes structural errors. We use Gemini3Pro for all generation.

Generation and Execution Pipeline. For each question, we sample $M = 3$ SQL queries per interpretation, yielding $3 \times 3 = 9$ generated queries per model-temperature configuration. SQL generation, entity resolution, self-correction, and execution follow Ziletti and DAmbrosi (2024). Queries are run on Optum's de-identified Clinformatics® Data Mart Database, and retrieved results are verbalized into natural language answers, from which we compute H_R, $\tilde{H}_{R|I}$, and $H_{R|I}$ as in Sec. 3. Evaluation

follows Sec. 4.1 with $\eta = 0.6$.

5.3. Results

Results are presented in Table 3. $H_{R|I}$ outperforms H_R as a failure predictor in 17 of 20 model-temperature configurations, achieving pooled AUROC of 0.762 [95% CI: 0.737, 0.787] versus 0.600 [0.575, 0.627] for H_R. This difference is significant ($p < 10^{-10}$), demonstrating that conditioning on interpretation structure provides substantial discriminative value. The naive subtraction $\tilde{H}_{R|I}$ achieves only 0.461 AUROC with 60% negative values, validating the theoretical necessity of the Schur complement (Sec. 3.3).

		AUROC			
Model (T)	**Acc.**	H_R	$\tilde{H}_{R\|I}$	$H_{R\|I}$	Δ
Qwen3					
T=0.7	53.3	.605	.544	**.764**	+54.3
T=1.0	54.2	.617	.531	**.748**	+42.6
T=1.5	55.0	.618	.505	**.818**	+59.2
T=2.0	54.1	.630	.498	**.786**	+65.8
GPT-OSS					
T=0.7	71.6	.600	.416	**.731**	+31.2
T=1.0	75.5	.615	.410	**.828**	+34.3
T=1.5	71.6	.615	.384	**.831**	+34.2
T=2.0	71.6	.612	.391	**.818**	+35.0
KimiK2					
T=0.7	76.1	.632	.446	**.815**	+41.7
T=1.0	74.3	.548	.417	**.817**	+40.5
T=1.5	75.2	.635	.444	**.791**	+44.7
T=2.0	78.9	.599	.418	**.853**	+27.8
Claude4.5S					
T=0.7	78.9	.593	.395	**.611**	+0.0
T=1.0	77.6	.595	.466	**.694**	+10.3
T=1.5	78.0	.633	.400	**.744**	+30.8
T=2.0	78.9	.575	.426	**.622**	+16.8
Gemini3Pro					
T=0.7	93.3	**.708**	.623	.604	+3.8
T=1.0	93.3	.620	.532	**.715**	+20.6
T=1.5	88.8	**.642**	.578	.593	+5.9
T=2.0	88.8	**.635**	.542	.595	+10.7
Pooled	*74.4*	*.600*	*.461*	***.762***	*+33.6*

Table 3: Clinical Text-to-SQL benchmark results ($N = 2{,}180$; 109 questions $\times$ 5 models $\times$ 4 temperatures). Δ = Regime A $-$ Regime B accuracy gap. $\tilde{H}_{R|I}$ produces negative values in 60% of cases.

Regime Analysis Confirms Discriminative Value. The accuracy gap Δ between Regime A (high H_R, low $H_{R|I}$) and Regime B (high H_R, high $H_{R|I}$), split at median values, is positive in 19 of 20 configurations (pooled: +33.6 pp). This replicates the open-domain QA pattern (Sec. 4.2): among queries with equally high output diversity, those with low $H_{R|I}$ are substantially more likely to succeed.

Temperature Modulates Uncertainty Signal. For most models, moderate temperatures ($T = 1.0$–1.5) yield the strongest $H_{R|I}$ signal. Claude 4.5 illustrates this pattern: regime separation increases from $\Delta = 0$ at T=0.7 to $\Delta = +30.8$ at T=1.5 ($0 \rightarrow +30.8$), suggesting higher sampling diversity exposes latent instability. Gemini 3 Pro is the exception: its errors appear to stem from consistent but incorrect outputs rather than instability.

Sensitivity to Diffusion Parameter τ. Table 4 shows $H_{R|I}$ is robust for $\tau \geq 2$ (AUROC 0.75–0.76), while H_R degrades at high τ (0.74 $\rightarrow$ 0.58) as diffusion blurs pairwise similarities toward uniformity. The Schur complement resists this effect: both raw similarities and the projection term smooth proportionally, preserving residual signal.

τ	1	2	3	5	10	15	20
H_I	.571	.574	.576	.578	.572	.569	.567
H_R	.669	.742	.738	.676	.601	.586	.582
$H_{R\|I}$	.675	.746	.764	.765	.762	.761	.760

Table 4: Pooled AUROC sensitivity to heat kernel parameter τ. $H_{R|I}$ remains stable across $\tau \in [2, 20]$ while H_R degrades at high diffusion.

6. Clinical Text-to-SQL in Deployment Settings

6.1. Experimental Setup

Having established that $H_{R|I}$ outperforms H_R with known interpretations, we now test whether these gains extend to production deployment. This setting differs from previous experiments in three key ways.

Interpretations are generated, not given. Interpretations are produced on the fly via our disambiguation procedure (Sec. 5.1) and may imperfectly capture the true ambiguity structure, unlike the validated gold/silver labels used previously.

Uncertainty serves as a proxy signal. We evaluate final-answer correctness only, meaning $H_{R|I}$ computed over interpretations serves as a proxy for reliability rather than directly measuring per-interpretation accuracy.

Limited statistical power. Evaluation is coarser (one label per question rather than per interpretation), and models achieve high accuracy (Table 5), yielding sparse errors that challenge statistical estimation.

We use the same pipeline as Sec. 5.2, testing with and without disambiguation and RAG. RAG

follows the methodology in Ziletti and DAmbrosi (2024).

6.2. Results

Table 5 summarizes performance in the deployment setting. All models achieve 84–100% accuracy, higher than Sec. 5.3. This reflects the evaluation setup: strict path consistency requires correctness across multiple interpretations including less natural ones, while here we evaluate a single answer against the original query, which typically matches the model's default interpretation.

Dis	RAG	Qwen3	GPT-OSS	KimiK2	Claude4.5S	Gemini3Pro	Pooled
×	×	85.3	93.6	99.1	95.4	100	94.7
✓	×	88.1	98.2	99.1	94.5	97.2	95.4
×	✓	90.8	99.1	97.2	97.2	100	96.9
✓	✓	94.5	92.7	94.5	96.3	100	95.6
Pooled		89.7	95.9	97.5	95.9	99.3	95.6

Table 5: Model Accuracy by Configuration. Dis. = disambiguated input. All values in %.

Uncertainty Separates Correct from Incorrect. Despite sparse errors, median uncertainty values are systematically higher for incorrect predictions. Pooled across all configurations ($N = 2{,}180$), incorrect predictions show median H_I of 0.189 vs. 0.069 for correct, H_R of 0.442 vs. 0.301, and $H_{R|I}$ of 0.072 vs. 0.003.

RAG Stabilizes Output. RAG substantially reduces $H_{R|I}$ for correct predictions: GPT-OSS drops from 0.113 to 0.030, Qwen3 from 0.046 to 0.006, and KimiK2 from 0.062 to 0.025. However, RAG also reduces $H_{R|I}$ for incorrect predictions (pooled median: 0.104 $\rightarrow$ 0.052), which may suppress the uncertainty signal that would otherwise flag errors (e.g., GPT-OSS: 0.296 $\rightarrow$ 0.025). Claude4.5S and Gemini3Pro show minimal $H_{R|I}$ (0.002) with or without RAG, suggesting sufficient internalized domain knowledge.

AUROC and regime-based diagnostics. In this setting (Sec. 6.1), pooled AUROC is 0.687 [0.629, 0.740] for H_R versus 0.648 [0.600, 0.701] for $H_{R|I}$ (not significant, $p = 0.20$). Crucially, H_R cannot distinguish whether a high-uncertainty query needs clarification, model review, or both. CLUES is designed to *decompose* uncertainty rather than maximize single-score prediction, providing diagnostic resolution unavailable from H_R alone (Sec. 6.3).

6.3. Uncertainty Regime Analysis

Given sparse per-condition errors, we pool all data for regime analysis. We partition queries using pooled median thresholds for H_I and $H_{R|I}$, yielding a 2×2 structure (Table 6).

Regime	H_I	$H_{R\|I}$	C/I	Err
I: Confident	Low	Low	534/8	1.5%
II: Ambiguity	High	Low	530/18	3.3%
III: Instability	Low	High	529/21	3.8%
IV: Compound	High	High	492/48	**8.9%**
Total			2085/95	4.4%

Table 6: Uncertainty Regime Analysis. 2×2 decomposition by H_I and $H_{R|I}$ median thresholds. C/I = Correct/Incorrect.

Error Gradient and Routing Implications. Error rates increase from Regime I (1.5%) to Regime IV (8.9%). Each regime contains ∼25% of queries, yet Regime IV concentrates 51% of all errors (48/95). A routing strategy sending only Regime IV to human review would examine one quarter of queries while catching half of all failures. Within-regime AUROC in Regime IV (≈0.52 for all metrics) indicates that continuous entropy values provide limited additional signal once the regime is determined. The decomposition's value is further demonstrated among high-uncertainty queries where both H_I and H_R are above median ($N = 712$): these appear identical under conventional metrics, yet $H_{R|I}$ separates them into Regime A (4.3% error) and Regime B (9.2% error). This difference is significant (odds ratio OR = 2.24, $p < 0.02$) and holds across all five models using a single global threshold.

7. Summary

We introduced CLUES, a framework that decomposes semantic uncertainty into ambiguity (H_I) and conditional instability ($H_{R|I}$) via a Schur-complement construction over an interpretation–result bipartite graph. Across open-domain QA and clinical Text-to-SQL, $H_{R|I}$ is competitive with, and often improves upon, state-of-the-art Kernel Language Entropy (Nikitin et al., 2024) for failure prediction, while providing diagnostic resolution unavailable from a single score. Regime analysis based on $(H_I, H_{R|I})$ surfaces distinct failure patterns and maps them to targeted interventions: clarification for high ambiguity, human review for high instability. CLUES is model-agnostic and requires only output sampling; future work will explore adaptive routing and uncertainty propagation in agentic pipelines.

8. Limitations

Computational Cost. CLUES requires multiple LLM calls per query (interpretations $\times$ samples; 9 in our setup). Calls can be parallelized, though cost remains a consideration for large-scale deployment.

Interpretation Quality. CLUES assumes generated interpretations meaningfully capture query ambiguity. If the disambiguation procedure produces trivial or redundant interpretations, H_I may underestimate true input ambiguity, and the Schur complement may not isolate instability effectively. We observed strong results with frontier LLMs, but interpretation quality likely degrades with weaker models.

Routing Efficiency in Sparse Error Regimes. In high-accuracy settings (95.6% in deployment), regime-based routing catches 51% of errors by reviewing 25% of queries, twice the efficiency of random selection, though not yet sufficient for fully automated triage. The fundamental challenge is sparse errors: with only 95 failures ($\sim$4% error rate), fine-grained uncertainty signals have limited discriminative power. Applications with higher base error rates may benefit more substantially from regime-based routing.

9. Bibliographical References

Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. 2024. Distinguishing the knowable from the unknowable with language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Anthropic AI. 2023. Model card and evaluations for claude models. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf. Accessed: February 15, 2024.

Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press, Cambridge, England.

Spiros Denaxas, Kenan Direk, Arturo Gonzalez-Izquierdo, Maria Pikoula, Aylin Cakiroglu, Jason Moore, Harry Hemingway, and Liam Smeeth. 2017. Methods for enhancing the reproducibility of biomedical research findings using electronic health records. *BioData Mining*, 10:31.

Naihao Deng, Yulong Chen, and Yue Zhang. 2022. Recent advances in text-to-SQL: A survey of what we have and what we expect. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2166–2187, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Darius Foodeei, Simin Fan, and Martin Jaggi. 2025. Semantic uncertainty in advanced decoding methods for llm generation. *arXiv preprint arXiv:2506.17296*.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. Technical report, Google. Accessed: February 15, 2024.

Shrestha Ghosh, Moritz Schneider, Carina Reinicke, and Carsten Eickhoff. 2025. A survey on LLM-assisted clinical trial recruitment. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 625–646, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Yashvir S. Grewal, Edwin V. Bonilla, and Thang D. Bui. 2024. Improving uncertainty quantification in large language models via semantic embeddings. *arXiv preprint arXiv:2410.22685*.

Sebastian G. Gruber and Florian Buettner. 2024. A bias-variance-covariance decomposition of kernel scores for generative models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Nicholas J. Higham. 2002. Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, 22(3):329–343.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

Yongrae Jo, Seongyun Lee, Minju Seo, Sung Ju Hwang, and Moontae Lee. 2024. LG AI research & KAIST at EHRSQL 2024: Self-training large language models with pseudo-labeled unanswerable questions for a reliable text-to-SQL system on EHRs. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 635–643, Mexico City, Mexico. Association for Computational Linguistics.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*.

Daeyoung Kim, Seongsu Bae, Seungho Kim, and Edward Choi. 2022. Uncertainty-aware text-to-program for question answering on structured electronic health records. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 138–151. PMLR.

Sangryul Kim, Donghee Han, and Sehyun Kim. 2024. ProbGate at EHRSQL 2024: Enhancing SQL query generation accuracy through probabilistic threshold filtering and error handling. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 687–696, Mexico City, Mexico. Association for Computational Linguistics.

Mathew J Koretsky, Maya Willey, et al. 2025. BiomedSQL: Text-to-SQL for scientific reasoning on biomedical knowledge bases. *arXiv preprint arXiv:2505.20321*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *International Conference on Learning Representations (ICLR)*.

Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong C Lee, and Edward Choi. 2023. EHRSQL: A practical text-to-SQL benchmark for electronic health records. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Xiaomin Li, Zhou Yu, Ziji Zhang, Yingying Zhuang, Swair Shah, Narayanan Sadagopan, and Anurag Beniwal. 2025a. Semantic volume: Quantifying and detecting both external and internal uncertainty in llms. *arXiv preprint arXiv:2502.21239*.

Zhaoye Li, Siyuan Shen, Wenjing Yang, Ruochun Jin, Huan Chen, Ligong Cao, and Jing Ren. 2025b. Enhancing uncertainty quantification in large language models through semantic graph density. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, UAI '25. JMLR.org.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: fine-grained uncertainty quantification for llms from semantic similarities. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

OHDSI Community. 2024. Atlas: Open source software for observational data analysis. https://github.com/OHDSI/Atlas. Accessed: 2026-01-28.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. 2025. gpt-oss-120b and gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Shalini Raghuvanshi, Yashs Tiwari, and Abhishek Yadav. 2025. Integrating token-level uncertainty, bidirectional nli, and semantic entropy for robust hallucination detection in large language models. In *2025 International Conference on Artificial intelligence and Emerging Technologies (ICAIET)*, pages 1–6.

Jaehee Ryu, Seonhee Cho, Gyubok Lee, and Edward Choi. 2024. EHR-SeqSQL : A sequential text-to-SQL dataset for interactively exploring electronic health records. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16388–16407, Bangkok, Thailand. Association for Computational Linguistics.

Irina Saparina and Mirella Lapata. 2024. Ambrosia: a benchmark for parsing ambiguous questions into database queries. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Irina Saparina and Mirella Lapata. 2025. Disambiguate first, parse later: Generating interpretations for ambiguity resolution in semantic parsing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16825–16839, Vienna, Austria. Association for Computational Linguistics.

J Schur. 1917. Über potenzreihen, die im innern des einheitskreises beschränkt sind. *J. Reine Angew. Math.*, 1917(147):205–232.

Haoyi Song, Ruihan Ji, Naichen Shi, Fan Lai, and Raed Al Kontar. 2025. Inv-entropy: A fully probabilistic framework for uncertainty quantification in language models. *arXiv preprint arXiv:2506.09684*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi Wang, Dinglu Wang, Feng Wang, Haiming Wang, Jianzhou Wang, Jiaxing Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Xiaofei Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang Yuan, Mengjie Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yangkun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian Zhao, Yikai Zhao, Huabin Zheng, Shaojie Zheng, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Fuzhen Zhang, editor. 2005. *The schur complement and its applications*, 2005 edition. Numerical Methods and Algorithms. Springer, New York, NY.

Michael Zhang and Eunsol Choi. 2021. Situat-

edQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angelo Ziletti and Leonardo DAmbrosi. 2024. Retrieval augmented text-to-SQL generation for epidemiological question answering using electronic health records. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 47–53, Mexico City, Mexico. Association for Computational Linguistics.

Angelo Ziletti and Leonardo D'Ambrosi. 2025. Generating patient cohorts from electronic health records using two-step retrieval-augmented text-to-sql generation. *arXiv preprint arXiv:2502.21107*.

Meredith N. Zozus, Rachel L. Richesson, Anita Walden, Jessie D. Tenenbaum, and W. E. Hammond. 2016. Research reproducibility in longitudinal multi-center studies using data from electronic health records. *AMIA Joint Summits on Translational Science Proceedings*, 2016:279–285.

An OMOP-Based Open-Source Text-to-SQL Benchmark Dataset

Paul Legrand[1], **Kawsar Noor**[1], **Satyam Bhagwanani**[1], **Richard J Dobson**[1,2]
[1] University College London, London, United Kingdom
[2] King's College London, London, United Kingdom
paul.legrand@etu.univ-amu.fr, kawsar.noor.15@ucl.ac.uk,
s.bhagwanani@ucl.ac.uk, richard.j.dobson@kcl.ac.uk

Abstract

Access to electronic health record (EHR) warehouses is limited by SQL expertise and complex clinical schemas. We present an open-source OMOP Common Data Model text-to-SQL benchmark (CDM v5.4) with a safety contract: output one executable SQL statement or the abstention token (<NO_SQL>) for unanswerable requests. Inputs are concept-normalized (entities as OMOP concept IDs) to decouple SQL generation from entity linking. We evaluate by executing predicted and reference queries on a synthetic OMOP PostgreSQL database, reporting Execution Accuracy (result equivalence) and a reliability score that rewards correct abstention and penalizes unsafe attempts. The dataset includes 6,690 paraphrases from 75 OMOP-adapted templates with leakage-resistant template/SQL-variation splits. LoRA-tuned Llama-3-8B-Instruct achieves 93.55% execution accuracy with improved abstention reliability, while schema-injected baselines fail the contract. We release the dataset, splits, database dump, and a reproducible evaluation pipeline to support reliable clinical analytics assistants.

Keywords: Text-to-SQL, Electronic Health Records (EHR), OMOP CDM, Large Language Models (LLMs)

1. Introduction

EHRs hold rich structured clinical data, yet extracting cohorts or summary statistics still requires SQL and schema expertise. Text-to-SQL could enable natural-language querying, but clinical deployment is difficult: performance is highly schema-dependent, and seemingly plausible SQL can silently yield incorrect cohorts. Evaluation should therefore prioritize executable, schema-faithful SQL and quantify risk-sensitive reliability, including abstention (Geifman and El-Yaniv, 2019) when a request is not answerable from the database. General benchmarks such as Spider (Yu et al., 2018) have driven progress, but their schemas and evaluation settings differ from real hospital warehouses, limiting transfer; execution-based scoring is essential because syntactically valid SQL can still be clinically wrong due to faulty joins or predicates. Earlier systems often relied on explicit schema linking and relation-aware schema encoders (e.g., RAT-SQL (Wang et al., 2020)).

Recent clinical datasets add templated questions and unanswerables to test abstention, yet benchmark-style resources explicitly grounded in OMOP-CDM remain scarce despite the model's widespread adoption. OMOP-oriented efforts are often narrower in scope (e.g., epidemiological question answering) and are seldom framed as engine-specific, fully executable benchmarks (e.g., PostgreSQL) with an explicit abstention/refusal contract. For instance, Ziletti and DAmbrosi (2024) explore OMOP-CDM text-to-SQL for epidemiological QA with a RAG-based pipeline; however, our setting differs in that we explicitly include non-answerable requests and specify a dedicated abstention token to support a refusal-aware executable benchmark.

The OMOP Common Data Model (CDM) is a standardized relational schema promoted by the Observational Health Data Sciences and Informatics (OHDSI) community (Hripcsak et al., 2015), validated for scalable observational safety surveillance across heterogeneous databases (Overhage et al., 2012), and evaluated as a practical CDM for longitudinal registries (Garza et al., 2016).

We cast clinician-style free-text requests as executable queries over OMOP CDM v5.4 under a strict interface: output either SQL only (one statement ending with exactly one semicolon) or the abstention token <NO_SQL>. Because OMOP concept names are unreliable keys, robust cohort definitions use concept_id-based predicates (optionally expanded via vocabulary relationships); we therefore encode medical constraints as concept identifiers to evaluate OMOP-grounded SQL generation rather than brittle string matching. Accordingly, this benchmark evaluates the downstream SQL generation component given concept-normalized inputs; integrating upstream clinical entity linking from raw text is left to future end-to-end work.

We release[1] an open-source benchmark and reproducible pipeline with:

- a text-to-SQL dataset grounded in OMOP CDM v5.4 with an explicit safe-abstention label (<NO_SQL>) for unanswerable requests;
- leakage-resistant train/validation/test splits that keep paraphrases grouped and control

[1] https://github.com/red1dwarf/TEXT-TO-OMOP

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 381–393
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

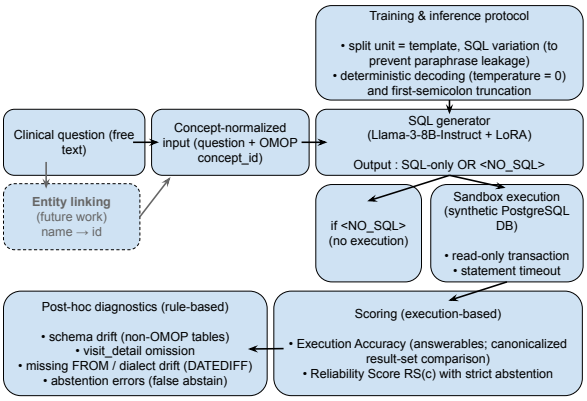


Figure 1: Overview of the benchmark and evaluation protocol.

for template/SQL variation overlap;

- a standardized evaluation suite combining text-based metrics, answerability detection, and execution-based scoring in read-only PostgreSQL, plus post-hoc diagnostic analyses;

- end-to-end reproducibility artifacts, including training/inference scripts (LoRA fine-tuning and schema-injected baselines) and an optional synthetic OMOP PostgreSQL dump.

2. Method

We describe the benchmark task, dataset construction, split protocol, training/evaluation setup, and metrics for OMOP CDM v5.4 text-to-SQL.

2.1. Task definition

We study schema-grounded SQL generation for EHR analytics under a strict executable contract. Given an input question x, the model must output either a single executable, OMOP-grounded SQL statement $\hat{q}$ (SQL only, no prose), ending with exactly one semicolon, or the exact abstention token `<NO_SQL>` when the question is unanswerable from the database (defined in Section 2.4). At inference, we enforce the single-statement constraint by truncating generation at the first semicolon. Any output that is not exactly `<NO_SQL>` is treated as a SQL attempt and executed for scoring; invalid or empty SQL thus results in an execution failure. All models are trained and evaluated with one instruction: use only OMOP CDM tables/columns and valid relationships; never output `SELECT *`; output only SQL ending with exactly one semicolon; otherwise abstain with `<NO_SQL>`. OMOP-grounding violations (e.g., schema drift) are detected via post-hoc diagnostics and reflected in execution-based metrics.

2.2. Target schema

The target environment is the OMOP Common Data Model (CDM) v5.4 using the PostgreSQL dialect (Observational Health Data Sciences and Informatics (OHDSI), 2021; PostgreSQL Global Development Group, 2025). Valid queries may reference only OMOP CDM tables/columns and must follow OMOP relationships and conventions (e.g., concept-driven filtering through standardized vocabularies and event-table semantics). Queries that reference non-OMOP schema elements are treated as schema drift and are penalized in diagnostic and execution-based evaluation.

2.3. Dataset construction

The benchmark includes 75 question templates aligned with OMOP join patterns and conventions. Starting from the EHRSQL template set (a template-based clinical text-to-SQL resource over EHR data; (Lee et al., 2023)), we built a local pool and retained a subset to limit structural redundancy (i.e., templates sharing the same query skeleton with only slot/value substitutions or minor surface differences). We then adapted these templates to OMOP CDM v5.4 join paths and conventions and added a small number of OMOP-specific templates to better cover under-represented domains and joins (e.g., `note`/`note_nlp` free-text tables, `provider` and `death`, and `episode`/`episode_event` patterns).

To capture free-text-linked querying, we included note-centric templates using `note` and `note_nlp`. In a local OMOP-mapped MIMIC-IV instance (Johnson et al., 2023; OHDSI, 2025), the `note` table was populated from the MIMIC-IV NOTE module and the `note_nlp` table from MedCAT-derived annotations over radiology reports and discharge summaries (Kraljevic et al., 2021). We used an OMOP-mapped MIMIC-IV environment to reflect common clinical warehouse settings where structured OMOP domains and unstructured text coexist; it also enabled us to check schema connectivity and template plausibility during development. This OMOP-mapped MIMIC-IV environment was used only during dataset/template development; execution-based evaluation is performed exclusively on the released synthetic OMOP PostgreSQL database (Section 2.8).

Each template is typically instantiated into three (template, SQL) variations: the SQL structure is fixed while inserted values (e.g., concept IDs, patient IDs, thresholds, time windows) vary, yielding 223 (template, SQL) variations.

Gold queries often use common table expressions (CTEs; `WITH`) to match clinical analytics practice and to make cohort definitions, temporal constraints, and visit-hierarchy logic explicit, encourag-

ing structured OMOP decompositions over monolithic nested subqueries.

For each variation, we produce 30 surface forms: one canonical question plus 29 GPT-4.1 paraphrases (OpenAI, 2025a). Paraphrases are constrained to (i) preserve meaning and (ii) keep placeholder values within typed tags unchanged, resulting in 6,690 answerable examples. Unanswerable questions are not included in this total; we take them from EHRSQL-2024 solely for abstention evaluation. The released *.jsonl files include these unanswerables alongside the answerable splits after applying the typed-tag and formatting normalization described in Section 2.4.

Typed identifier tags. Identifiers are enclosed in lightweight typed tags (e.g., <DRUG>...</DRUG>, <PERSON_ID>...</PERSON_ID>). These markers have no XML semantics; they enable concept-normalized inputs by representing clinical mentions as `concept_id` values (rather than free-text mentions or `concept_name` strings) while making the identifier type explicit. For instance, "Did patient 289 take Panadol yesterday?" becomes "Did patient <PERSON_ID>289</PERSON_ID> take <DRUG>839681</DRUG> yesterday?". Typing both concept identifiers (drugs, conditions, etc.) and non-concept identifiers (patient IDs) reduces ambiguity, discourages brittle associations between raw integers and OMOP fields, and encourages schema-driven reasoning (mapping typed roles to the correct *_concept_id predicates and join paths rather than relying on lexical cues).

Human verification protocol. Authors reviewed each paraphrase to ensure semantic equivalence, preservation of all slot values and temporal constraints, and absence of clinically implausible additions; failures were edited or removed.

2.4. Normalization of unanswerables

We use unanswerable questions from the EHRSQL-2024 shared task (Lee et al., 2024b). As in EHRSQL (Lee et al., 2023), a question is labeled unanswerable when the structured OMOP instance alone is insufficient: it either (i) requires external/normative clinical knowledge not stored in the database, or (ii) asks for information outside the database/schema.

In the initial setup, there was a shortcut: unanswerables often lacked typed tags, making tag presence predictive of answerability. In the released normalized splits (*.jsonl), we mitigate this by (i) inserting typed tags when entity-like identifiers occur, (ii) sampling synthetic identifiers to mirror answerable distributions (e.g., person-like ID lengths), and (iii) standardizing punctuation, thereby weakening the "tags $\Rightarrow$ answerable" cue.

2.5. Concept-normalized inputs and anti-memorization design

OMOP queries typically require standardized vocabulary identifiers (e.g., *_concept_id). Because mapping surface mentions (e.g., "troponin") to OMOP IDs is an upstream entity-linking step, each example includes two aligned inputs: `question_name` (human-readable concept names) and `question_concept` (key entities replaced by concept IDs). Importantly, there is no explicit name-to-ID linkage between the two views: the surface concept names in `question_name` do not correspond to the concept IDs appearing in `question_concept`. We release both for readability and future end-to-end studies, but report results only for SQL generation from the concept-normalized input (Figure 1), so metrics reflect generation with gold-like OMOP identifiers rather than raw-text entity linking. To limit spurious "ID binding", identifiers are synthetic (concept-/person-like integers) and not tied to real trajectories, though a minority may correspond to public OMOP concept_ids from the OHDSI Standardized vocabularies (Reich et al., 2024).

2.6. Unanswerables & abstention

Unanswerables are included to stress-test reliability. They include both normative/external-knowledge requests and "near-miss" out-of-schema queries (e.g., cross-ontology code lookups such as ICD-11$\rightarrow$ICD-10), mirroring realistic user intents that a database-backed assistant should refuse safely rather than answer with a plausible proxy SQL. For these examples, the only correct output is the single abstention token <NO_SQL> (no explanatory text), making abstention machine-actionable for downstream routing/clarification.

2.7. Split protocol

Instance-level splits can leak in paraphrase-heavy datasets, so we split at the (template, SQL) variation level: paraphrases of the same underlying query never cross splits. Per template, up to two variations are designated *seen*; each has 30 paraphrases split 20/5/5 into train/validation/test. All remaining variations are *unseen*; we sample 37 unseen variations for validation and 37 for test, adding 5 paraphrases per unseen variation to evaluate controlled generalization. Unanswerables are injected at fixed rates (10% train; 20% validation/test), yielding 3,312 train examples (2,980 answerable + 332 unanswerable) and 1,162 examples for both validation and test (930 + 232). The split is deterministic with a fixed seed (default 42), and we release the JSONL split files for exact reproducibility.

Category	Statistic (answerables)
Templates / variations	75 templates; 223 (template, SQL) variations
Size (answerable)	6,690 examples
JOINs (gold SQL)	median 1; p90 3; max 4
CTE usage (WITH)	40.36%
GROUP BY / DISTINCT	28.25% / 20.63%
ORDER BY / LIMIT	60.54% / 60.54%
Tagging (questions)	median 1 tag; 12.57% no-tag
Most-used OMOP tables	concept (45.74%), visit_occurrence (30.04%), condition_occurrence (26.91%), drug_exposure (21.52%), measurement (17.49%), visit_detail (16.14%), person (13.45%), procedure_occurrence (10.76%), device_exposure (10.76%)

Table 1: Dataset summary statistics for answerable examples.

2.8. Synthetic OMOP database for execution-based evaluation

Execution-based evaluation is the most faithful measure of text-to-SQL correctness (Wang et al., 2018), but with synthetic identifiers even correct SQL can return empty outputs on a real OMOP instance. We therefore release a controlled synthetic OMOP PostgreSQL database for evaluation only (never training). It is generated *post hoc* from the fixed gold SQL set (model-agnostic) and is not iteratively altered using model outputs: for each gold query, we insert the minimal synthetic rows needed in the relevant OMOP tables so that the gold query returns a non-empty result, while preserving basic heterogeneity (multiple patients, dates, events). Population was performed through GPT-5.1-assisted PostgreSQL insert scripts (OpenAI, 2025b), followed by author review, verification, and execution; OMOP integrity constraints were not enforced (duplicates may exist). All gold queries are verified to execute successfully, and we release the full database dump and profiling artifacts used for all reported execution metrics. Gold SQL queries were authored/validated against a real OMOP-mapped MIMIC-IV environment during dataset development (Section 2.3) and finalized before creating this synthetic instance, ensuring no circularity between query design and the evaluation database.

Profiling shows the instance is non-trivial: 23 OMOP tables with 117,010 rows (∼10 MB), including 1,858 persons and 6,453 visits. Event density is strongly long-tailed, with 125.48 events/person on average, a median of 9 events/person, a p99 of 992.92 events/person, and a maximum of 2,374 events/person. The visit hierarchy is substantial (63,826 `visit_detail` rows, ∼9.9× `visit_occurrence`), making join paths and aggregation/temporal logic challenging rather than "toy-like". The database is designed for discriminative scoring rather than epidemiological realism; for time-relative templates (e.g., "today"), we use a single fixed reference date (see Section 6).

2.9. Models and training setup

We fine-tune Llama-3-8B-Instruct (Grattafiori et al., 2024) with LoRA (Hu et al., 2021). Inputs follow the Llama-3 chat template with a fixed system prompt enforcing SQL-only outputs and <NO_SQL> abstention; loss is computed only on generated tokens (prompt masked). LoRA: $r = 32$, $\alpha = 64$, dropout 0.05, applied to attention and MLP projection modules. Training: 3 epochs, batch 8, grad accumulation 4, LR 1.5×10^{-4} with cosine schedule, warmup ratio 0.05, optionally using bf16 and/or gradient checkpointing.

2.10. Inference protocol

All models use deterministic inference (seed 42; temperature 0.0; top_p 1.0; max_new_tokens 1024), producing exactly one output per example (SQL or <NO_SQL>). Outputs are saved as JSONL and scored with the same execution-based pipeline on the released synthetic OMOP PostgreSQL instance. We score raw generations truncated at the first semicolon, with no model-specific post-processing.

Schema-injected baseline prompting. For baselines, we prepend an OMOP PostgreSQL DDL schema block (OMOP tables/columns only) to the user message for every example, aligning with Natural-SQL-style prompting. We run Natural-SQL-7B (ChatDB, 2024) (a public 7B ChatDB text-to-SQL instruction model), plus Llama-3-8B-Instruct (base) (Grattafiori et al., 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), under the same schema-injection setup. Although prompts follow each model's chat/template conventions, we keep an identical system instruction and the same SQL-only / <NO_SQL> output contract across all models (baselines and LoRA); the

injected schema block is the only prompt content that differs between baselines and LoRAs.

2.11. Evaluation metrics

We prioritize semantic correctness and reliability, and treat Execution Accuracy (ExecAcc) and the risk-sensitive Reliability Score (RS(c)) as the primary benchmark metrics for EHR text-to-SQL. All other metrics are reported as secondary analyses and diagnostics.

Primary text-to-SQL metrics.

Execution Accuracy (ExecAcc). Predicted and gold SQL are executed on PostgreSQL over the synthetic OMOP database (read-only transactions; per-query statement timeout). ExecAcc is computed only on answerable examples: a prediction is correct iff it is not <NO_SQL>, executes successfully, and returns the same normalized (order-invariant) result set as the gold query. Unanswerable examples are excluded from ExecAcc and evaluated via RS(c) and answerability-oriented metrics.

Formal definition. Let A be the set of answerable examples (gold $\neq$ <NO_SQL>) whose gold SQL executes without error on the evaluation database. Let $\hat{q}_i$ be the model output and $\mathrm{Res}(q)$ the normalized (order-invariant) result set returned by PostgreSQL. We define $\mathrm{Res}(q) = \perp$ if execution fails (e.g., syntax error, timeout), so the equality $\mathrm{Res}(\hat{q}_i) = \mathrm{Res}(q_i)$ can only hold for successful executions. ExecAcc is computed as:

$$\mathrm{ExecAcc} = \frac{1}{|A|} \sum_{i \in A} \mathbf{1}[\hat{q}_i \neq \texttt{<NO_SQL>} \wedge \mathrm{Res}(\hat{q}_i) = \mathrm{Res}(q_i)] \quad (1)$$

We compare gold and predicted results after normalization: (i) map NULL to the empty string, (ii) keep values in native Python types (no string casting), (iii) canonicalize each row as the sorted tuple of its values, sorted deterministically by grouping values by Python type, then by a stable textual sorting key, so the comparison is invariant to output-column permutations, and (iv) sort the resulting list of rows to be order-invariant.
Column names and column order are deliberately ignored. Mapping NULL to the empty string improves robustness to driver-specific NULL handling; in our evaluation set, gold queries for answerable questions never return NULL or empty strings, so this choice does not change the reported execution metrics.

Reliability Score (RS(c)). Following the risk-sensitive reliability evaluation proposed in TrustSQL (Lee et al., 2024a), we assign a sample-level score based on five cases, with penalty parameter $c \in \{0, 1, 10\}$:

1. answerable + correct SQL $\rightarrow +1$
2. answerable + abstain (<NO_SQL>) $\rightarrow 0$
3. answerable + incorrect/failed SQL $\rightarrow -c$
4. unanswerable + abstain (<NO_SQL>) $\rightarrow +1$
5. unanswerable + SQL attempt $\rightarrow -c$

Dataset-level RS(c) is the average sample score; higher c penalizes unsafe attempts more strongly than abstention, matching the clinical preference that wrong is worse than refuse. Reporting $c \in \{0, 1, 10\}$ shows how conclusions change from a neutral setting ($c = 0$), where unsafe SQL attempts are not penalized more than abstention, to increasingly risk-averse settings.

Secondary metrics (supporting analysis).

We also report: (i) Answerability F1 ($F1_{ans}$) for the binary choice SQL vs. <NO_SQL> (answer vs. abstain), with "answerable" as the positive class (SQL = positive, <NO_SQL> = negative); (ii) Unanswerable Recall (UnansRec), the fraction of unanswerables predicted as <NO_SQL>; (iii) False abstain (%), the fraction of answerables predicted as <NO_SQL>; and (iv) $F1_{exe}$, which summarizes executable correctness over attempted SQL outputs (with R_{exe} = ExecAcc by definition), computed as the harmonic mean of P_{exe} and R_{exe}. $F1_{exe}$ and $F1_{ans}$ are inspired by EHRSQL (Lee et al., 2023).

Formalization of $F1_{exe}$. Let A be answerable examples whose gold SQL runs without error on the evaluation database; let $T \subseteq A$ be those where the model outputs SQL ($\hat{q}_i \neq$ <NO_SQL>), and $C \subseteq T$ those executed correctly ($\mathrm{Res}(\hat{q}_i) = \mathrm{Res}(q_i)$). Unlike EHRSQL, we compute $\mathrm{Precision}_{exe}$ over T (conditioned on gold-answerables), since attempts on unanswerables are handled by our abstention/reliability metrics. Thus $\mathrm{Precision}_{exe} = |C|/|T|$, $\mathrm{Recall}_{exe} = |C|/|A|$ (= ExecAcc), and $F1_{exe} = 2PR/(P+R)$ with $P = \mathrm{Precision}_{exe}$ and $R = \mathrm{Recall}_{exe}$.

Auxiliary string-overlap and rule-based diagnostics. We report Exact Match (EM), Exact Match Normalized (EMN), and ROUGE-L as string-overlap diagnostics, but treat them as secondary because semantically equivalent SQL can differ syntactically (e.g., join order, predicate reordering) (Finegan-Dollak et al., 2018). EM is exact string match (Rajpurkar et al., 2016); EMN applies a lightweight normalization (lowercasing, trailing semicolon removal, whitespace collapsing);

ROUGE-L is based on the longest common subsequence (LCS) overlap (Lin, 2004). We also report rule-based diagnostics (e.g., schema drift, table-set mismatch, dialect drift, visit hierarchy omissions) to characterize common failure modes.

3. Results

We report results on the held-out test split under a unified pipeline: identical instruction (SQL-only; single statement; abstain with <NO_SQL>), deterministic decoding, truncation at the first semicolon, and scoring using execution-based and reliability-oriented metrics on the synthetic database.

3.1. Main results

Table 2 shows a clear gap between OMOP-adapted and general-purpose models. For transparency about training dynamics, we also report LoRA-150, which is not a separate model family but an earlier checkpoint from the same LoRA training run. The best adapter, LoRA-312, reaches 93.55% ExecAcc and $F1_{exe}$=93.65%, while preserving near-perfect answerability ($F1_{ans}$=99.46%) and positive risk-sensitive reliability (RS(10)=0.373). LoRA-150 and LoRA-312 should be read as an early-vs.-final checkpoint comparison within one training run, rather than as two independently trained methods.

Sensitivity analysis. With a stricter equivalence (matching column headers and preserving within-row value order, still row-order invariant), LoRA-312 performance remains essentially unchanged (93.44% ExecAcc), suggesting ExecAcc is not an artifact of result-set canonicalization (i.e., our default comparison ignores column names/order and canonicalizes rows; Section 2.11). Moreover, the released synthetic database enforces unique identifiers (distinct entities never share an ID) and gold answerable queries never return NULL/empty strings, so ambiguous identical-value multisets across columns are not expected in practice.

Off-the-shelf baselines remain far lower in executable correctness despite schema-injected prompting: Natural-SQL-7B (18.39%), Llama-3-8B-Instruct (base) (10.54%), and Mistral-7B-Instruct-v0.3 (4.09%). Their RS(10) is strongly negative (-8.381 to -9.517), showing that under our strict contract they frequently attempt SQL instead of abstaining, which is heavily penalized by reliability-oriented scoring. All baselines have EM/EMN = 0.00, largely due to incorrect SQL, while also suggesting that the rare correct outputs seldom match the gold SQL form, reinforcing the need for execution-based evaluation.

Overall, strict OMOP querying requires schema- and domain-aligned supervision beyond generic instruction or text-to-SQL tuning.

3.2. Answerability: abstention behavior

With 20% unanswerables in the test set (232/1,162), abstention must be evaluated explicitly. Under class imbalance (930 answerable vs 232 unanswerable), $F1_{ans}$ (SQL vs <NO_SQL>) can be inflated by models that almost always attempt SQL: baselines still reach $F1_{ans}$ ≈89% yet have RS(10)≈ -9, indicating unsafe over-attempting rather than calibrated abstention. We therefore prioritize Unanswerable Recall and RS(c) for safety-critical behavior. Table 3 shows that LoRA-312 abstains correctly on 96.55% (224/232) of unanswerables while rarely over-refusing answerables (0.22%, 2/930); LoRA-150 is slightly lower on unanswerables (92.67%) with the same false-abstain rate (0.22%). Baselines almost never abstain (UnansRec 0–5.60%), yielding large high-penalty reliability losses.

3.3. Diagnostics: failures & challenges

To complement aggregate scores, Table 3 reports interpretable failure proxies. False Abstain and Unanswerable Recall follow Section 3.2; Schema drift / Table-set mismatch / DATEDIFF usage / CTE usage are computed on gold-answerable items where the model attempts SQL (excluding <NO_SQL>), while Unanswerable Recall is computed on unanswerables. Schema drift flags any FROM/JOIN relation token that is neither an OMOP table, a CTE name, nor a whitelisted set-returning function. Table-set mismatch compares the set of base tables in gold vs. prediction (CTEs excluded). DATEDIFF usage detects PostgreSQL-incompatible DATEDIFF. CTE usage checks whether the prediction starts with WITH. visit_detail miss is computed on examples whose gold SQL uses visit_detail, counting a miss whenever the prediction omits visit_detail (or abstains).

Baselines show large table-set mismatch (52-79%) and frequent visit_detail omission (70-89%) when visit hierarchy reasoning is required, plus schema drift (non-OMOP tables in FROM/JOIN; up to 14.62% for Llama-3-8B-Instruct (base)) and dialect drift (e.g., DATEDIFF, ≈5-6% for some baselines). LoRA largely eliminates schema/dialect drift (≈0%) and reduces table-set mismatch to 0.65-1.19%, leaving visit hierarchy grounding as the key residual error: omitting visit_detail (1.33% for LoRA-312) can still produce executable SQL with clinically different semantics (e.g., ICU segment time vs. overall hospitalization).

3.4. Qualitative error cases (LoRA-312)

We highlight three representative residual failure modes for the best adapter (Table 4) and link each example to the corresponding diagnostic in Table 3.

Model	ExecAcc (%)	F1_exe (%)	F1_ans (%)	EM (%)	EMN (%)	ROUGE-L (%)	RS(0)	RS(1)	RS(10)
LoRA-312	**93.55**	**93.65**	**99.46**	**90.54**	**90.86**	**98.77**	**0.941**	**0.885**	**0.373**
LoRA-150	89.68	89.77	98.99	81.40	82.26	97.65	0.903	0.807	-0.052
Natural-SQL-7B	18.39	18.39	88.91	0.00	0.00	40.23	0.147	-0.706	-8.381
Llama-3-8B-Instruct (base)	10.54	10.54	89.21	0.00	0.00	36.74	0.090	-0.819	-9.006
Mistral-7B-Instruct-v0.3	4.09	4.09	89.47	0.00	0.00	31.08	0.043	-0.912	-9.517

Table 2: Model performance on the OMOP text-to-SQL benchmark (test set). LoRA-150 and LoRA-312 are two checkpoints from the same LoRA training run (early vs. final checkpoint). RS(0), RS(1), and RS(10) denote the Reliability Score with penalty parameter $c \in \{0, 1, 10\}$, where larger c penalizes unsafe SQL attempts more strongly relative to abstention.

Model	False abstain (%)	Unanswerable Recall (%)	Schema drift (%)	Table-set mismatch (%)	DATEDIFF usage (%)	CTE usage (%)	visit_detail miss (%)
LoRA-312	**0.22**	**96.55**	**0.00**	**0.65**	**0.00**	**40.30**	1.33
LoRA-150	0.22	92.67	0.00	1.19	0.00	40.09	**0.67**
Natural-SQL-7B	0.00	0.00	0.43	52.47	0.00	15.05	78.67
Llama-3-8B-Instruct (base)	0.00	3.02	14.62	76.88	5.81	0.32	70.00
Mistral-7B-Instruct-v0.3	0.00	5.60	8.17	79.25	5.27	9.68	88.67

Table 3: Diagnostic analysis on the test set (930 answerable / 232 unanswerable).

4. Discussion

Results show that OMOP-specific adaptation is necessary: generic SQL skill does not transfer, and remaining errors are concentrated in join-path, visit-hierarchy, and temporal/aggregation semantics.

4.1. Why generic SQL fails in OMOP

We observe a large gap between general instruction-tuned baselines and the OMOP-adapted LoRA model: baselines often produce plausible "SQL-like" outputs that fail in OMOP, where queries must follow domain conventions (concept-driven filtering via standardized vocabularies, event-table semantics, and canonical join paths across person, visits, and clinical-domain tables). Diagnostic results make this explicit: baseline errors are dominated by table-set mismatch, visit-hierarchy mistakes, and schema/dialect drift, showing that generic SQL competence does not yield OMOP-correct queries. Conversely, the best LoRA adapter suggests these conventions can be learned with modest parameter-efficient tuning in a lightweight open-source model, enabling on-prem hospital deployment and CDM/dialect adaptation without relying on large or proprietary systems.

4.2. Reliability and abstention should be first-class evaluation targets

Clinical analytics is unlike many text-to-SQL settings: a wrong query can silently produce plausible but incorrect cohorts or statistics. Accordingly, the benchmark includes unanswerable questions and enforces a single abstention token, `<NO_SQL>`. Reliability-aware scoring (RS(c)) complements ExecAcc by increasingly penalizing unsafe attempts relative to abstention as c grows, reflecting a common clinical preference: wrong is worse than refuse. The LoRA model's higher RS(c), together with high unanswerable recall and low false-abstain rates, shows abstention is both learnable and quantifiable under a simple, machine-actionable policy. Practically, free-form explanations inside the SQL-generation component may be counterproductive: "helpful" natural language can be ambiguous for downstream systems and mistaken for executable intent. A strict contract (SQL-only or `<NO_SQL>`) better supports safe agentic integration, where `<NO_SQL>` can trigger clarification, fallback tools, or human review.

4.3. Remaining challenges: joins, visits, time/aggregation

After fine-tuning, Table 3 shows the remaining gap is semantic, concentrated in three OMOP-critical choices: (i) join-path selection across event tables when multiple schema-consistent routes exist (especially for multi-domain queries); (ii) visit-hierarchy selection (`visit_occurrence` vs. `visit_detail`), where choosing the wrong level or omitting required `visit_detail` implicitly changes the cohort; and (iii) temporal/aggregation semantics (window bounds,

Failure mode	Representative example (gold vs. LoRA-312)
Visit hierarchy grounding (`visit_detail` omission; aligns with *visit_detail miss* in Table 3)	**Question:** How long has patient <PERSON_ID>6102133621015</PERSON_ID> been in the ICU during the current hospital stay? **Gold (uses ICU segment time):** `SELECT ... FROM visit_occurrence vo JOIN visit_detail vd ON vd.visit_occurrence_id = vo.visit_occurrence_id JOIN concept c ON vd.visit_detail_concept_id = c.concept_id WHERE ... c.concept_name ILIKE '%intensive care%' ...;` **LoRA-312 (drops `visit_detail`):** `SELECT ... FROM visit_occurrence vo JOIN concept c ON vo.visit_concept_id = c.concept_id WHERE ... c.concept_name ILIKE '%intensive care%' ...;` **Impact:** executable SQL but clinically different semantics (hospital-level stay vs. ICU segment duration).
Table grounding confusion (`measurement` vs. visit tables; aligns with *Table-set mismatch* in Table 3)	**Question:** Can you find the third instance of patient <PERSON_ID>3720242571871</PERSON_ID>'s weight recorded on May 6, last year? **Gold (OMOP measurement retrieval + offset):** `SELECT m.value_as_number FROM measurement m WHERE m.person_id = 3720242571871 AND m.measurement_concept_id = 3025315 AND m.measurement_date = MAKE_DATE(EXTRACT(YEAR FROM CURRENT_DATE)::int - 1, 5, 6) ORDER BY m.measurement_datetime OFFSET 2 LIMIT 1;` **LoRA-312 (incorrect table choice):** `SELECT ... FROM visit_occurrence vo JOIN concept c ON ... WHERE vo.person_id = 3720242571871 AND vo.visit_start_date = MAKE_DATE(...) AND c.concept_name ILIKE '%weight%' ORDER BY ... OFFSET 2 LIMIT 1;` **Impact:** wrong OMOP event-table semantics; misses concept-driven filtering and measurement ordering logic.
Unsafe attempt on an unanswerable (aligns with *Unanswerable Recall* and RS(c) penalties in Table 3)	**Question:** How long should patient <PERSON_ID>4773708619089828657</PERSON_ID> have taken <DRUG>942384</DRUG>, the last month? **Gold:** <NO_SQL> **LoRA-312 (unsafe SQL attempt):** `SELECT MAX(de.drug_exposure_start_datetime) ..., MIN(...) ... FROM drug_exposure de WHERE de.person_id = ... AND de.drug_concept_id = 942384 AND EXTRACT(MONTH FROM de.drug_exposure_start_date) = EXTRACT(MONTH FROM CURRENT_DATE) - 1;` **Impact:** answers a normative/external-knowledge request by (incorrectly) fabricating a computable proxy from timestamps.

Table 4: Representative residual failure modes for LoRA-312.

grouping granularity, `DISTINCT`), where small mismatches can look plausible yet invalidate the statistic. This implies (a) the benchmark remains discriminative for clinically consequential reasoning, including schema-grounded relational structure, event- vs. visit-level intent, and faithful temporal/aggregation cues, and (b) progress likely requires structured grounding and semantic verification beyond scaling alone. Three directions are: (1) constraint-aware decoding over an OMOP join graph (e.g., PICARD; Scholak et al. (2021)) to down-weight or block invalid table transitions; (2) selective schema retrieval that returns only the relevant tables, keys, and concept mappings, reducing distractors from full-schema injection (e.g., RASL; Eben et al. (2025)); and (3) targeted post-generation checks of high-impact commitments, including alignment of time predicates with the question's temporal framing and of aggregation level (including `DISTINCT`) with the requested measure (e.g., Ren et al. (2025)). Overall, the residual gap is schema-grounded semantic faithfulness, motivating hybrid generate-and-check pipelines for reliable clinical text-to-SQL.

5. Conclusion

In conclusion, this benchmark provides an OMOP CDM v5.4-grounded, reliability-oriented testbed for EHR text-to-SQL under a strict, deployment-friendly output contract. The results show that parameter-efficient adaptation can achieve high executable correctness and substantially improved abstention behavior, while general-purpose baselines remain brittle under OMOP constraints. We expect this resource to support research on safer clinical analytics assistants that combine upstream entity linking with reliable SQL generation, constrained execution, and robust abstention/routing strategies. By providing an OMOP-grounded, execution-scored benchmark with calibrated abstention, we aim to accelerate safer free-text interfaces for cohort discovery and clinical analytics.

6. Limitations

We report results on concept-normalized inputs (entities expressed as OMOP concept identifiers), so we do not evaluate end-to-end entity linking from raw clinical text (e.g., using a clinical entity linker such as MedCAT (Kraljevic et al., 2021)). While we release both human-readable and concept-normalized questions to support end-to-end work, real deployments will face concept ambiguity and linking errors that can propagate to SQL execution. In deployment, linking uncertainty may require confirmation or abstention policies (e.g., abstaining when concept ambiguity is high).

Our split protocol is designed to test controlled generalization to unseen (template, SQL-variation) structures and paraphrased surface forms. It does not directly measure cross-institutional distribution shift across independent OMOP deployments, where data distributions, local conventions, and integrity constraints may differ; absolute execution accuracy may therefore change on real institutional OMOP databases, even if the strict SQL-only/<NO_SQL> output contract and our reliability-oriented evaluation remain applicable.

The benchmark is also limited in coverage and lexical diversity by its template-based construction; future work should expand diversity with closer coordination to real hospital information needs.

In addition, our unanswerable examples are adapted from EHRSQL-2024 rather than constructed natively for OMOP-specific failure modes, so future work should curate OMOP-native unanswerables tailored to schema-specific abstention scenarios.

Finally, execution-based scoring relies on a controlled synthetic OMOP PostgreSQL instance populated so that gold queries return non-empty outputs, enabling discriminative execution scoring while preserving the anti-memorization design. This synthetic database is not intended to mirror real prevalence, coding distributions, or integrity constraints; absolute execution metrics may therefore differ in practice. For time-sensitive templates (e.g., "today"), we fix the reference date to 2026-02-02 in the released evaluation dump.

7. Ethics Statements

This work supports clinical data access through a strict text-to-SQL (or <NO_SQL>) interface and is intended for research use. The released benchmark includes (i) natural-language questions, (ii) gold SQL queries, and (iii) a synthetic OMOP CDM v5.4 PostgreSQL dump designed for discriminative execution-based evaluation.

No MIMIC-IV content is distributed: we do not release any MIMIC-derived clinical notes, patient-level records, or other protected/identifiable health information.

Despite the strict output contract and execution-based evaluation, model-generated SQL may still be incorrect or unsafe if used without safeguards; we recommend defense-in-depth measures (e.g., sandboxed execution, timeouts, query allow/deny-lists, logging, and human oversight for high-stakes use).

8. Supplementary Materials

8.1. Dataset construction

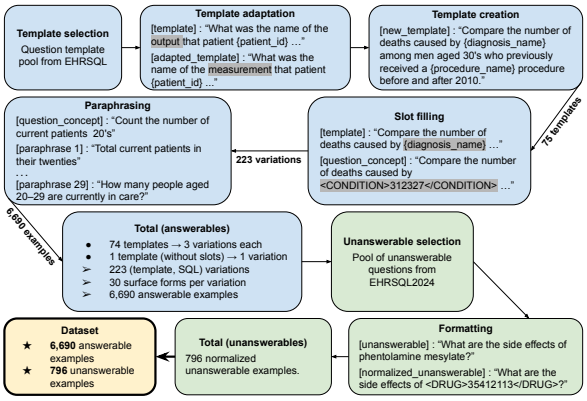


Figure 2: Dataset construction overview. OMOP-adapted EHRSQL templates yield 223 (template, SQL) variations and 6,690 answerable examples; separately selected and normalized EHRSQL-2024 unanswerables contribute 796 additional examples.

8.2. Split protocol

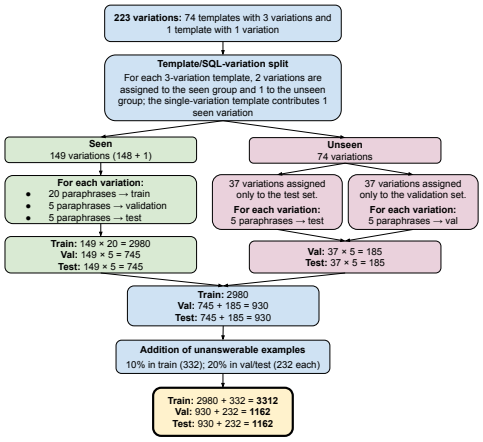


Figure 3: Leakage-resistant split construction at the (template, SQL)-variation level. Seen variations contribute to train/validation/test, unseen variations only to validation or test, and unanswerables are added at fixed rates.

8.3. Brief description of OMOP tables used in benchmark-specific examples

To improve accessibility for readers less familiar with OMOP CDM, Table 5 briefly summarizes several benchmark-relevant OMOP tables that may be less intuitive from their names alone.

Table	Role in OMOP / relevance here
visit_occurrence	Stores the coarse-grained care encounter (e.g., hospitalization or outpatient visit). Used for visit-level reasoning and last-visit queries.
visit_detail	Stores finer-grained sub-visit segments within a visit (e.g., ICU stay within a hospitalization). Important when the clinically correct answer depends on sub-visit granularity.
note	Stores unstructured clinical notes as text documents linked to patients and visits.
note_nlp	Stores NLP-derived annotations over clinical notes (e.g., extracted concepts/mentions), enabling note-linked structured querying.
episode	Represents a higher-level clinical episode grouping related events over time.
episode_event	Links an episode to the constituent OMOP events that belong to it.

Table 5: Short glossary of OMOP tables that may be less familiar to non-OMOP readers but appear in benchmark construction or error analysis.

8.4. Concrete examples of released instances

To make the benchmark format more explicit, we provide below one answerable example and one unanswerable example from the released JSON instances. As described in the main text, each answerable item may include a human-readable question (question_name), a concept-normalized question (question_concept), the gold SQL query, and the originating template. Unanswerable examples use the same input format but the only correct target output is the abstention token <NO_SQL>.

Answerable example. This example illustrates a concept-normalized input where the drug mention is replaced by an OMOP concept_id, while the corresponding human-readable version is also retained for readability.


```
{
  "question_name": "What are the top 4
      frequent prescribed drugs for
      patients who were also prescribed
      <DRUG>bisacodyl 10 MG Rectal
      Suppository [Bisac-Evac]</DRUG> at
       the same time on the last
      hospital visit?",
  "question_concept": "What are the top
      4 frequent prescribed drugs for
      patients who were also prescribed
      <DRUG>19041823</DRUG> at the same
      time on the last hospital visit?",
  "sql": "WITH last_visits AS (SELECT
      DISTINCT ON (vo.person_id) vo.
      visit_occurrence_id FROM
      visit_occurrence AS vo ORDER BY vo
      .person_id, vo.visit_start_date
      DESC), patients_with_19041823 AS (
      SELECT lv.visit_occurrence_id FROM
       last_visits AS lv JOIN
      drug_exposure AS de1 ON lv.
      visit_occurrence_id = de1.
      visit_occurrence_id WHERE de1.
      drug_concept_id = 19041823),
      co_prescribed_drugs AS (SELECT de2
      .drug_concept_id FROM
      drug_exposure de2 JOIN
      patients_with_19041823 AS p ON de2
      .visit_occurrence_id = p.
      visit_occurrence_id WHERE de2.
      drug_concept_id != 19041823)
      SELECT d.drug_concept_id AS
      concept_id, c.concept_name, COUNT
      (*) AS count FROM
      co_prescribed_drugs AS d JOIN
      concept AS c ON d.drug_concept_id
      = c.concept_id GROUP BY d.
      drug_concept_id, c.concept_name
      ORDER BY count DESC LIMIT 4;",
  "template": "what are_verb the top [
      n_rank] frequent prescribed drugs
      for patients who were also
```

```
      prescribed {drug_name} [
      time_filter_within] [
      time_filter_global1]?"
}
```

This example highlights the benchmark design choice of evaluating SQL generation from concept-normalized inputs: the model is expected to map the typed identifier <DRUG>19041823</DRUG> to the correct OMOP predicate structure (here, drug_exposure.drug_concept_id) and produce a single executable PostgreSQL query.

Unanswerable example. The following example illustrates the abstention setting. The question asks for a drug side effect, which cannot be answered from the OMOP database instance alone under our task definition. The only correct output is therefore the abstention token <NO_SQL>.

```
{
  "input": "What is the side effect of <
      DRUG>43833975</DRUG>?",
  "output": "<NO_SQL>"
}
```

This example belongs to the unanswerable portion of the benchmark and reflects the intended safety contract of the task: if the answer requires external or normative medical knowledge rather than information available in the structured OMOP database, the model must abstain rather than generate a plausible but unsupported SQL query.

8.5. Inference prompts used for LoRA models and schema-injected baselines

For reproducibility, we report below the exact prompt structure used at inference time. All models were run with deterministic decoding (temperature=0.0, top_p=1.0), and generations were truncated at the first semicolon so that outputs satisfied the single-statement contract. The same core instruction was used across LoRA models and baselines, with the main difference that baseline prompts additionally injected the PostgreSQL DDL schema into the user prompt.

Shared system instruction. The following system message was used in inference:

```
You are an assistant that generates
    valid PostgreSQL SQL queries for
    data stored in an electronic health
    record following the OMOP Common
    Data Model (CDM). Follow these rules
     strictly:
- Use only OMOP CDM tables, columns, and
     relationships.
- Never invent tables or columns.
- Never hallucinate schema details.
- Never use SELECT *.
- Terminate the SQL query with a
    semicolon.
- Output only the SQL query without
    prose, comments, or explanations.
- If the question cannot be answered
    from OMOP CDM data, output <NO_SQL>.
- Stop after the first semicolon.
```

LoRA inference prompt (Llama-3 chat format). For LoRA inference, the model was prompted in native Llama-3 chat format without schema injection. The user message only contained the question:

```
<|begin_of_text|><|start_header_id|>
    system<|end_header_id|>
[SYSTEM_MSG]<|eot_id|>
<|start_header_id|>user<|end_header_id|>
Generate the SQL query that answers the
    following question:
[QUESTION]
<|eot_id|>
<|start_header_id|>assistant<|
    end_header_id|>
```

Concretely, [SYSTEM_MSG] is replaced by the shared system instruction above, and [QUESTION] is replaced by the benchmark input string.

Schema-injected baseline prompts. For baseline models, the schema was injected into the prompt through the user message. The shared user-side content was:

```
You are given the database schema in
    PostgreSQL DDL.
Use it to write a correct query.

### Database schema (DDL)
[SCHEMA_DDL]

### Question
[QUESTION]
```

Here, [SCHEMA_DDL] denotes the OMOP PostgreSQL DDL block and [QUESTION] denotes the benchmark input.

Llama-3 baseline prompt. The Llama-3 baseline used the same native chat wrapper as the LoRA model, but with the schema-injected user block:

```
<|begin_of_text|><|start_header_id|>
    system<|end_header_id|>
[SYSTEM_MSG]<|eot_id|>
<|start_header_id|>user<|end_header_id|>
[USER_BLOCK_WITH_SCHEMA]
<|eot_id|>
<|start_header_id|>assistant<|
    end_header_id|>
```

where [USER_BLOCK_WITH_SCHEMA] is the schema-injected content shown above.

Mistral baseline prompt. The Mistral baseline used the tokenizer chat template when available, with the same two-message structure: (i) the shared system instruction, and (ii) the schema-injected user message. When the tokenizer chat template was unavailable, the fallback plain-text prompt was:

```
System: [SYSTEM_MSG]

User: [USER_BLOCK_WITH_SCHEMA]

Assistant:
```

This preserves exactly the same instruction content while adapting formatting to the tokenizer implementation.

Natural-SQL baseline prompt. The Natural-SQL baseline used the following task-style prompt:

```
# Task
[SYSTEM_MSG]

Generate a SQL query to answer the
    following question: `[QUESTION]`

### PostgreSQL Database Schema
The query will run on a database with
    the following schema:

[SCHEMA_DDL]

# SQL
Here is the SQL query that answers the
    question: `[QUESTION]`
```sql
```

Thus, the same instruction constraints were retained, but embedded in the Natural-SQL prompt format.

9. References

ChatDB. 2024. Natural-sql-7b. Hugging Face model card. Accessed: 2026-01-31.

Jeffrey Eben, Aitzaz Ahmad, and Stephen Lau. 2025. Rasl: Retrieval augmented schema linking for massive database text-to-sql. *arXiv preprint arXiv:2507.23104*. Accepted to the KDD Workshop on Structured Knowledge for Large Language Models (SKnowLLM '25).

Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-SQL evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 351–360, Melbourne, Australia. Association for Computational Linguistics.

Maryam Garza, Guilherme Del Fiol, Jessica Tenenbaum, Anita Walden, and Meredith Nahm Zozus. 2016. Evaluating common data models for use with a longitudinal community registry. *Journal of Biomedical Informatics*, 64:333–341.

Yonatan Geifman and Ran El-Yaniv. 2019. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR.

Aaron Grattafiori, Abhimanyu Dubey, Anish Jha, et al. 2024. The llama 3 herd of models.

George Hripcsak, Jon D. Duke, Nigam H. Shah, Christian G. Reich, Vojtech Huser, Martijn J. Schuemie, Marc A. Suchard, Rae Woong Park, Ian C. K. Wong, Peter R. Rijnbeek, Johan van der Lei, Nicole Pratt, G. Niklas Norén, Yu-Chuan (Jack) Li, Paul E. Stang, David Madigan, and Patrick B. Ryan. 2015. Observational health data sciences and informatics (ohdsi): Opportunities for observational researchers. In *Studies in Health Technology and Informatics*, volume 216, pages 574–578.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

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):1.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D. Shah, Wai Keong Wong, Zina Ibrahim, James T. Teo, and Richard
```

J. B. Dobson. 2021. Multi-domain clinical natural language processing with medcat: The medical concept annotation toolkit. *Artificial Intelligence in Medicine*, 117:102083.

Gyubok Lee, Woosog Chay, Seonhee Cho, and Edward Choi. 2024a. Trustsql: A reliability benchmark for text-to-sql models with diverse unanswerable questions. *arXiv preprint arXiv:2403.15879*.

Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. 2023. Ehrsql: A practical text-to-sql benchmark for electronic health records. *arXiv preprint arXiv:2301.07695*.

Gyubok Lee, Sunjun Kweon, Seongsu Bae, and Edward Choi. 2024b. 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*, pages 644–654, Mexico City, Mexico. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Observational Health Data Sciences and Informatics (OHDSI). 2021. Omop common data model (cdm) v5.4 specifications. Specification webpage. Accessed: 2026-01-31.

OHDSI. 2025. Ohdsi/mimic: Mimic iv to omop cdm conversion (etl pipeline). GitHub repository (release v1.0.0). Accessed: 2026-02-11.

OpenAI. 2025a. Gpt-4.1 model (openai api documentation). Developer documentation. Accessed: 2026-02-11.

OpenAI. 2025b. Gpt-5.1 instant and gpt-5.1 thinking system card addendum. System card addendum (PDF). Accessed: 2026-02-11.

J. Marc Overhage, Patrick B. Ryan, Christian G. Reich, Abraham G. Hartzema, and Paul E. Stang. 2012. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*, 19(1):54–60.

PostgreSQL Global Development Group. 2025. Postgresql documentation (current release). Online documentation. Accessed: 2026-01-31.

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, Austin, Texas. Association for Computational Linguistics.

Christian Reich, Anna Ostropolets, Patrick Ryan, Peter Rijnbeek, Martijn Schuemie, Alexander Davydov, Dmitry Dymshyts, and George Hripcsak. 2024. Ohdsi standardized vocabularies—a large-scale centralized reference ontology for international data harmonization. *Journal of the American Medical Informatics Association*, 31(3):583–590.

Tonghui Ren, Chen Ke, Yuankai Fan, Yinan Jing, Zhenying He, Kai Zhang, and X. Sean Wang. 2025. The power of constraints in natural language to sql translation. *Proceedings of the VLDB Endowment*, 18(7):2097–2111.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. Picard: Parsing incrementally for constrained auto-regressive decoding from language models. *arXiv preprint arXiv:2109.05093*.

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2020. Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7567–7578.

Chenglong Wang, Kedar Tatwawadi, Marc Brockschmidt, Po-Sen Huang, Yi Mao, Oleksandr Polozov, and Rishabh Singh. 2018. Robust text-to-sql generation with execution-guided decoding.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir R. Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3911–3921.

Angelo Ziletti and Leonardo DAmbrosi. 2024. Retrieval augmented text-to-SQL generation for epidemiological question answering using electronic health records. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 47–53, Mexico City, Mexico. Association for Computational Linguistics.

Profiling Hallucinations in Frontier LLMs for Entity Linking to Medical Ontologies

Logan Born[1], **Nishant Kambhatla**[1], **Uliyana Kubasova**[1], **Maryam Siahbani**[1,2], **Andrei Vacariu**[1], **Tim O'Connell**[1] **and Anoop Sarkar**[1]
[1]Emtelligent [2]University of the Fraser Valley
{lborn,nishant,uliyana,maryam,andrei,tim,anoop}@emtelligent.com

Abstract

The integration of Large Language Models (LLMs) into healthcare promises to revolutionize clinical documentation and interoperability, yet reliability remains a concern. This study presents a comprehensive analysis of hallucinations by frontier LLMs tasked with mapping clinical text to SNOMED CT. Through rigorous experimentation, we identify a critical reliability gap: LLMs hallucinate medical codes at a rate that currently renders them unsuitable for autonomous clinical coding applications. Paradoxically, constraining models to use ground-truth mention spans exacerbates, rather than mitigates, these hallucinations. We further contribute a taxonomy of hallucination types – including deprecated codes and cross-ontology errors – and demonstrate that general-purpose LLMs significantly underperform compared to specialized zero-shot entity linking approaches. These findings underscore the need for robust verification mechanisms before clinical deployment.

Keywords: clinical nlp, snomed-ct, ontologies, entity linking, LLMs, hallucination

1. Introduction

The transition from specialized discriminative models to general-purpose LLMs has prompted an ongoing re-evaluation of clinical NLP pipelines (Shool et al., 2025; Maity and Saikia, 2025). While frontier LLM providers such as Anthropic[1] and OpenAI[2] advocate for LLMs as end-to-end agents in healthcare settings, the overall utility of these models depends heavily on their ability to ground their output into established knowledge bases; in clinical settings, this grounding may be achieved through entity linking to an ontology such as SNOMED CT.

However, treating medical entity linking as a generative task introduces a novel failure mode absent in discriminative baselines: structural hallucination (Huang et al., 2025). Unlike traditional classification models that select from a fixed candidate set, generative models construct identifiers token-by-token. This creates the possibility of generating plausible-looking but non-existent identifiers. Evaluating these hallucinations in open-domain text is notoriously difficult due to the ambiguity between factual errors and legitimate inference (ibid.).

Medical entity linking against an ontology such as SNOMED CT offers a uniquely tractable testbed for quantifying LLM faithfulness by facilitating hallucination detection. SNOMED CT Identifiers (SCTIDs) are not arbitrary strings: they belong to a strictly controlled vocabulary with internal structure and built-in validation features.[3]These properties allow us to disentangle *alignment errors*, where the model retrieves an incorrect but valid concept, from *hallucinations*, where a model fabricates a code that is structurally invalid or non-existent.

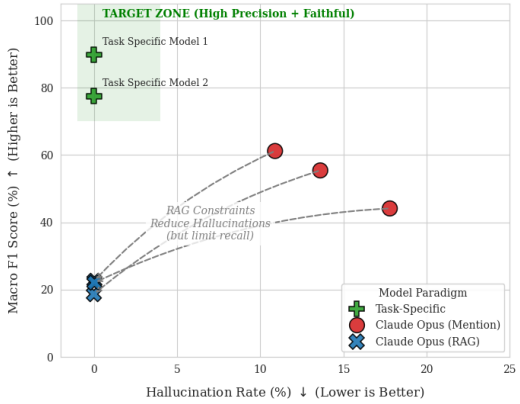


Figure 1: Comparison of generalist and specialist models on a clinical entity linking task. Task-specific architectures like KIRI (Davidson et al., 2025) achieve high extraction performance without hallucinating. In contrast, Generalist LLMs exhibit a strict trade-off: they either suffer from high hallucination rates in end-to-end settings or suffer from low recall in RAG settings, failing to simultaneously optimize for both faithfulness and accuracy.

We investigate the capacity of frontier LLMs to perform mention-based linking against SNOMED CT, comparing them against state-of-the-art discriminative architectures. Our experiments reveal

[1] https://www.anthropic.com/news/healthcare-life-sciences

[2] https://openai.com/index/openai-for-healthcare/

[3] https://docs.snomed.org/snomed-ct-specifications/snomed-ct-release-file-specification/snomed-ct-identifiers/6.4-check-digit

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 394–413
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

significant shortcomings in LLMs' ontological reasoning abilities (Figure 1):

High Hallucination Rates: In standard prompting setups, frontier LLMs fail to respect the closed vocabulary requirement, with 20–55% of generated codes being invalid.

Precision-Recall Tradeoff: Retrieval-Augmented Generation (RAG) mitigates hallucination, but also severely bottlenecks recall. Generalist models struggle to synthesize retrieved information and fail to match task-specific model performance.

Complex Failure Modes: We profile errors from generative models, which include the mixing of disparate ontology extensions (e.g., UK vs. US editions) and the resurrection of deprecated codes.

Ultimately, our results suggest that medical coding remains a challenging domain where specialized, discriminative architectures significantly outperform generalist generative approaches in both safety and accuracy.

2. Methodology

We propose to prompt three top-of-the-line large language models (Claude Sonnet 4.5, Claude Opus 4.6, and GPT-5.2) to perform entity linking tasks on a corpus of discharge summaries. We aim first to quantify how frequently these models hallucinate in their output: of the concept IDs which they produce, what proportion really exist in the US release of the SNOMED CT ontology? Secondly, we aim to describe qualitatively *how* these models hallucinate: what proportion of their outputs are outdated, invalid, from nonstandard ontology extensions, or are pure fabrications.

We aim to show that, despite advances in grounding LLM outputs through techniques like retrieval-augmented generation, even frontier models remain prone to fabricating outputs. While some of these hallucinations are immediately obvious, and would likely be caught and pruned through a simple post-hoc filtering step, others are more subtle and may not be immediately evident as errors.

3. Experimental Setup

3.1. Data

We use a revised version of the SNOMED CT Entity Linking Challenge dataset (Hardman et al., 2025), which comprises 272 discharge summaries from MIMIC-IV (Johnson et al., 2023).

We manually update the annotations in the original dataset using the 2025-03-01 US Edition of SNOMED CT, which combines the International Release with the US Extension. Additionally, we re-annotate these documents using a commercial medical NLP engine (Appendix A). The original dataset labels some phrases inconsistently, such that (for example) `BLOOD Glucose` is occasionally labelled as `Glucose measurement, blood` and identical strings in other settings are unlabelled (see Davidson et al. 2025 for additional discussion). The original dataset also intentionally excludes some SNOMED CT sub-hierarchies. Our reannotation efforts are intended to make the dataset more consistent and to provide coverage for more parts of the SNOMED CT ontology.

We prompt Claude Haiku[4] 4.5 to act as an independent judge, filtering the NLP engine outputs by flagging any that are incorrect (either inherently, or due to some feature of the surrounding context), irrelevant, or too broad or narrow. For example, in "`pt was initially oriented x1`", "`oriented`" is assigned the concept `Oriented to person, time and place (finding)`, when the context implies only a single degree of orientation; the LLM therefore filters this annotation from the final test set. We take the union of any surviving annotations with the annotations from the original dataset; new annotations take priority in case of overlap.

The resulting dataset comprises the same 272 documents as the original (68 test, 204 train) and a total of 78961 annotated links to SNOMED concepts (24087 test annotations, 54874 train annotations). It covers 6271 distinct SNOMED concepts (5151 in training, 3845 in test; 1120 concepts are zero-shot, occurring only in the training set), with a minimum frequency of 1, maximum of 786, and median of 2.[5]

This reannotation effort clearly biases any entity linking evaluation in favour of our own model. However, the intended focus of this work is on analyzing and taxonomizing hallucinations; we are not primarily interested in entity linking accuracy *per se*, and we therefore consider this an acceptable bias. We include our model in the tables of results to show the expected upper bound of performance achievable when there is perfect alignment between the test set and a model's level of detail and phrase segmentation choices, but grey it out to highlight that it is not a fair comparand.

3.2. Baselines

To rigorously evaluate the capabilities of frontier LLMs in clinical concept normalization and to isolate the sources of hallucination, we devised four distinct experimental configurations. These set-

[4]We use a small model for this task as we believe this type of pass/fail judgment does not require the full reasoning power of a larger model.

[5]Upon request, we will share a patch allowing parties with access to the original SNOMED CT Entity Linking Challenge dataset to update to our revised version.

tings range from isolating the linking capability to evaluating full pipeline performance and testing retrieval-augmented strategies.

Oracle Span In this setting, we decouple entity recognition from entity linking. The LLM is provided with the clinical note text alongside the ground-truth character offsets for target entities. The model is tasked with predicting the correct SNOMED CT concept ID for each provided span. To manage context window limitations and optimize throughput, entities are processed in batches (50 per inference call). This setting serves as an upper bound for the model's internal parametric knowledge, isolating its ability to map clinical language to ontology codes without the confounding effects of extraction errors.

End-to-End (E2E) This is the most stringent evaluation, simulating a raw deployment scenario. The LLM is provided only with the raw clinical text and must perform joint Named Entity Recognition and Normalization (NER+N). The model is tasked with extracting entity mentions, determining their exact character start and end offsets, and predicting the associated SNOMED CT code. While this tests the viability of a monolithic LLM pipeline, it penalizes models for tokenization artifacts; LLMs relying on sub-word tokenizers often struggle with precise character counting, resulting in scoring failures despite semantically correct extraction.

Mention-based Extraction (Mention) To mitigate errors from imprecise character counting while still evaluating extraction capabilities, we attempt a Mention-based approach. The LLM is instructed to extract entities by providing the surface text and occurrence count, for example, `("hypertension", 3)` denotes the third instance of the string "`hypertension`". We resolve these predictions to specific spans via post-hoc programmatic resolution using case-insensitive string matching. This setting allows us to assess the model's ability to identify and normalize clinical concepts based on semantic context, independent of its ability to perform arithmetic character indexing.

Retrieval Augmented Generation (RAG) We additionally implement a RAG pipeline that moves the burden of knowledge retrieval from the LLM's parametric memory to an external index. This follows a three-step pipeline:

1. **Extraction** The LLM extracts mentions (surface form and occurrence index) from the raw text, but does not predict any concept IDs.
2. **Retrieval** Extracted mentions are converted into embedding vectors and queried against a database of pre-computed embeddings of SNOMED CT Fully Specified Names (FSNs). We retrieve the top 5 candidate concepts based on cosine similarity. OpenAI's `text-embedding-3-small` is used to build 512 dimensional embeddings [6].
3. **Selection** The LLM is presented with the extracted mention and the retrieved candidate list, and must select the most appropriate concept ID. By constraining the LLM's output space to retrieved candidates, this setting aims to minimize hallucinations of non-existent or irrelevant medical codes.

We evaluate three state-of-the-art generalist models: Claude Opus 4.6, Claude Sonnet 4.5, and GPT-5.2. To ensure a robust comparative analysis, we employ a full factorial experimental design, evaluating every model across all four settings (Oracle Span, E2E, Mention, and RAG). This comprehensive evaluation allows us to directly compare the impact of architectural differences against methodological constraints, disentangling the models' intrinsic reasoning capabilities from the difficulty of the extraction and retrieval tasks.

To put the results in perspective, we compare against a closed-source commercial entity linking system called *entity-paradigm*, against AWS Medical Comprehend[7], and against KIRI, an open-source dictionary-based system which was the winning submission to the original SNOMED CT entity linking challenge (Davidson et al., 2025). These models do not generate codes from outside the target ontology, and therefore serve as a hallucination-free baseline. We retrain KIRI on the updated training annotations; the other comparands are zero-shot and see only the test set.

We report micro- and macro-averaged hallucination rates for all settings, which we define as the proportion of concept IDs in the output which do not exist in the 2025-03-01 US release of SNOMED CT, which was the current release at the time of these experiments.

We report macro-averaged character IoU using the official scoring script (Davidson et al., 2025) in keeping with prior work on the SNOMED CT Entity Linking Challenge dataset. We additionally report precision, recall, and macro-F1 for settings where the model must extract mentions prior to

[6] While the hyperparameter space for RAG (e.g., retrieval depth k, re-ranking strategies) is extensive, we fix $k = 5$ for all experiments. Given the significant length of clinical notes, a smaller k minimizes context dilution and ensures the LLM remains focused on the most relevant candidates. Furthermore, our primary objective is to evaluate the architecture's impact on hallucination mitigation rather than to optimize retrieval performance.

[7] https://aws.amazon.com/comprehend/medical/

linking (discriminative model, E2E LLM, and the LLM Mention setting). For the LLM Oracle Span setting, where the model is provided the spans, we instead report the accuracy of the assigned concept links.

4. Results

4.1. Hallucination Rate

In settings without RAG, the micro-averaged hallucination rate, i.e. the overall proportion of concept ID tokens which do not represent a real SNOMED concept, ranges from a minimum of 10.89% in the Claude Opus (Mention) setting to a maximum of 19.34% from Claude Sonnet (Oracle Span).

The macro-averaged hallucination rate, i.e. the proportion of *distinct* concept ids which are not real SCTIDs, is much higher, between 19.47% in the GPT-5.2 (End-to-end) setting and 52.50% in the Claude Sonnet (Oracle Span) setting.

As may be expected, RAG significantly reduces the hallucination rate for all models, though notably it does not eliminate hallucinations entirely: both Claude Opus and Sonnet hallucinate a single SCTID even in the RAG setting.

4.2. Character IoU and F1

As seen in Tables 1 and 2, generalist LLMs fail to achieve competitive macro-averaged character IoU scores when compared to task-specific models.

The end-to-end approach using GPT-5.2 performs particularly poorly, achieving just 0.015 IoU. This setting requires the model to not only produce accurate concept labels, but also to accurately count the character spans associated with these labels, a challenging task for the LLMs.

Scores in the Mention setting, where the model is required to perform entity extraction and linking, are better, though still do not exceed 8.98%.

The Oracle Span setting, which tests models only on their ability to assign correct concept labels, produces the highest IoU of any LLM-based approach, though even this score of 12.36% falls well short of the scores achieved by task-specific methods.

In both the E2E and Mention settings, precision generally outpaces recall, implying that all models regardless of provider fail to extract a majority of possible mentions. Recall is particularly poor in the E2E setting, where more than 90% of mentions go undiscovered, but even the best-case (Claude Opus, Mention) misses nearly half of all mentions.

Only for the metric of precision do LLM-based techniques begin to approach dictionary-based methods, where the best precision (72.25%, achieved by GPT-5.2, Mention) is just 5 points short of the prior state-of-the-art on this dataset.

Unlike the other models in this comparison, note that AWS Medical Comprehend outputs a top-k list of candidate concepts. Table 1 reports top-5 scores for this model. We note that the top-1 prediction is often semantically valid but differs in scope from the ground truth, e.g. `Adverse reaction` when the desired target is `Adverse reaction to drug`. A top-1 evaluation penalizes these cases, leading to a very low F1 score of just 0.25%. Performing a top-5 evaluation reduces the penalty for near-misses of this sort, as the broader or narrower target concept is often present in the list of outputs, albeit with a lower score.

5. Analysis

5.1. Hallucination Rate

The asymmetry between micro- and macro-averaged hallucination rates implies that models do not consistently hallucinate the same codes many times. Rather, there are a small number of codes which are hallucinated frequently, followed by a wide range of codes attested a few times apiece. This is visualized in Figure 2 (top), which shows the very long tail of rare hallucinations produced in the GPT-5.2 (Mention) setting. Note that frequency alone is not a sufficient condition for identifying hallucinations: real concept ids follow a similarly long-tailed distribution (Figure 2, bottom).

5.2. Impact of Frequency

There is, at best, a very weak correlation between concept frequency and a model's tendency to hallucinate. We estimate the overall frequencies of SNOMED CT concepts by annotating 100k discharge summaries using a commercial medical NLP engine and counting the number of occurrences of every SCTID in the resulting output. We measure Pearson's ρ between a concept's frequency in this dataset and the proportion of outputs that were hallucinated in the Oracle Span setting when the target label equalled the concept in question. We find $\rho = -0.0386$, $p = 0.0007$, i.e. there is a significant but extremely weak tendency for rarer concepts to induce more hallucinations.

There is no significant correlation ($\rho = -0.00642$, $p = 0.2308$) between the frequency of a mention and the likelihood that a model will hallucinate the concept assigned to that mention. In other words, LLMs appear equally likely to hallucinate codes for one-off abbreviations, misspellings, and rare wordings as for more frequent spans of text.

5.3. Entity extraction and hallucinations

In the Mention and E2E settings, there is the possibility that an LLM will extract a mention for which no

Task-Specific Models						
Model	**IoU ↑**	**P ↑**	**R ↑**	**F1 ↑**	**H ↓**	**Macro H ↓**
emtelligent entity-paradigm (Ours)	78.06%	91.95%	87.84%	89.85%	0.00%	0.00%
KIRI (Davidson et al., 2025)	50.49%	77.33%	77.91%	77.61%	0.00%	0.00%
Amazon Comprehend Medical†	22.06%	40.22%	49.91%	22.37%	0.00%	0.00%
General-Purpose Frontier LLMs						
Model	**IoU ↑**	**P ↑**	**R ↑**	**F1 ↑**	**H ↓**	**Macro H ↓**
GPT-5.2 (E2E)	0.15%	58.79%	9.46%	16.30%	12.21%	19.47%
Claude Opus 4.6 (E2E)	2.23%	2.34%	2.95%	2.61%	13.61%	42.21%
Claude Sonnet 4.5 (E2E)	1.60%	1.56%	2.15%	1.81%	18.84%	55.04%
GPT-5.2 (Mention)	5.39%	**72.25%**	44.97%	55.44%	13.61%	28.31%
Claude Opus 4.6 (Mention)	8.98%	68.48%	**55.33%**	**61.21%**	10.89%	24.94%
Claude Sonnet 4.5 (Mention)	4.78%	72.20%	31.79%	44.14%	17.79%	34.53%
GPT 5.2 (RAG)	11.66%	15.59%	23.27%	18.67%	**0.00%**	**0.00%**
Claude Opus 4.6 (RAG)	**14.56%**	20.37%	25.47%	22.64%	0.003%	0.02%
Claude Sonnet 4.5 (RAG)	14.08%	19.62%	25.19%	22.06%	0.003%	0.02%

Table 1: Macro-averaged Intersection over Union (IoU), precision (P), recall (R), F1 (F1), hallucination rate (H), and macro-averaged hallucination rate (Macro H) on SNOMED CT Entity Linking Challenge; KIRI is the 2025 winning submission (Davidson et al., 2025). †AWS Comprehend score uses top-5 outputs. Others use top-1 output.

Model	**IoU ↑**	**H ↓**	**Macro H ↓**	**Acc (Real) ↑**	**Acc (Overall) ↑**
emtelligent entity-paradigm (Ours)	84.45%	0%	0%	87.84%	87.84%
GPT-5.2 (Oracle Span)	6.69%	14.28%	38.19%	34.79%	29.83%
Claude Opus 4.6 (Oracle Span)	12.36%	13.57%	34.47%	52.43%	45.31%
Claude Sonnet 4.5 (Oracle Span)	8.02%	19.34%	52.50%	48.11%	38.81%

Table 2: Macro-averaged Intersection over Union (IoU), hallucination rate (H), macro-averaged hallucination rate (Macro H), accuracy limited to SCTIDs which actually exist (Acc (Real Only)) and overall accuracy (Acc (Overall)) on SNOMED CT Entity Linking Challenge when models are provided oracle access to the ground truth span boundaries.

SCTID exists, i.e. it may extract spans which should not and cannot be labelled using the target ontology. In the Oracle Span setting, this cannot happen unless the ground truth annotations themselves are incorrect. A reasonable hypothesis would be that LLMs are more likely to hallucinate codes for these spans, since there is no valid label to assign and since LLMs struggle with refusing to produce output in such cases (Pan et al., 2025).

For all models, we count which mention texts are associated with the highest frequency of hallucinated codes. For the 100 mentions with the highest hallucination rates, we check whether the ground truth data contains any instance where an identical mention was assigned to any concept. If such an instance exists, this implies that there is an SCTID which the model *could* have found to avoid hallucinating; if no such instance exists, it implies that the given mention may not actually correspond to any SCTID, in which case hallucination may have been inevitable. Table 3 reports how many of these mentions exist in the ground truth data.

Across all models, at least half of the highly-hallucinatory mention texts *do* exist in the ground truth. In other words, hallucinated SCTIDs are not simply a consequence of poor entity extraction.

Table 4 shows the top-5 spans of text which are most likely to be associated with a hallucinated concept ID for each model and setting. We observe a remarkable consistency in the set of terms associated with the most hallucinations: a majority are very short abbreviations, many relating to blood panel measurements. Although the same *terms* are associated with a high degree of hallucination across all models, the hallucinations themselves vary. Claude Opus and Sonnet tend to fabricate codes that resemble an SCTID, or to include SCTIDs which are real but outdated. GPT-5.2 has a stronger tendency to ignore the prompt telling it to produce SNOMED CT codes, and instead produces codes from other vocabularies such as LOINC. These codes are occasionally valid for the text in question, e.g. `787-2` for `MCV` with meaning `MCV [Entitic mean volume] in Red Blood Cells by Automated count`; though just as often they are incorrect, such as `33728-7` meaning `Size.maximum dimension in Tumor` assigned to the same text.

Model	
GPT-5.2 (E2E)	50%
GPT-5.2 (Mention)	75%
GPT-5.2 (Oracle Span)	100%
Claude Opus 4.6 (E2E)	79%
Claude Opus 4.6 (Mention)	66%
Claude Opus 4.6 (Oracle Span)	100%
Claude Sonnet 4.5 (E2E)	81%
Claude Sonnet 4.5 (Mention)	72%
Claude Sonnet 4.5 (Oracle Span)	100%

Table 3: Of the 100 mentions which receive the most hallucinations, what proportion exist in the ground truth data. A low percentage implies that a model hallucinates primarily in cases where there is no valid code to assign, i.e. when it has no choice but to either hallucinate or refuse to produce output. A higher percentage implies that a model is prone to hallucinate even in cases where there is a valid code which it could output.

5.4. SNOMED Extensions

Across settings, all of the LLM baselines and Amazon Medical Comprehend output a mixture of codes from multiple SNOMED extensions in addition to codes from the "standard" International edition.

GPT-5.2 includes codes from 5, 7, and 8 different extensions in the End-to-End, Mention, and Oracle Span settings respectively; Opus includes 7, 5, and 2; and Sonnet includes 10, 6, and 5. (This ignores invalid extensions such as `3000119`, which does not exist in the SNOMED CT Namespace Identifier Registry; across all models we observe roughly a dozen such nonexistent extensions in total.) AWS Comprehend outputs codes from 14 distinct extensions spanning organizations from four continents, although all of these are real extensions rather than invented ones.

Looking at the Oracle Span setting, where Opus includes just 2 distinct extensions, we find that these come from namespace `1000000`, belonging to the United Kingdom's National Health Service, and namespace `1000119`, belonging to Kaiser Permanente, an American corporation. These extensions therefore reflect vastly different healthcare systems, national versus corporate levels of organization, and consequently different needs and motivations underlying the choice of concepts included in their respective extensions. Thus, while these codes are not incorrect *per se*, they nevertheless represent a degree of inconsistency which may negatively impact the usefulness of these model outputs in a healthcare setting. This inconsistency will be all the greater in settings which mix codes from a larger number of extensions.

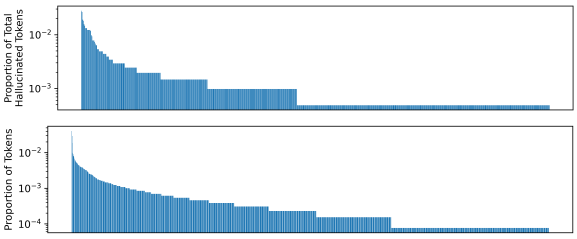


Figure 2: Rate of repetition of hallucinated codes (top) and real codes (bottom) in the GPT-5.2 (Mention) setting; each bar represents one concept id. Other settings exhibit a similarly long tail of rare and one-off codes.

5.5. Deprecated Codes

Table 5 reports the proportion of codes from each model which are listed as "inactive" in the United States edition 2025-03-01 release of SNOMED.

Between 3% and 5% of codes are deprecated on average, except in RAG settings where less than 1% of codes are deprecated. The proportion of deprecated codes is otherwise roughly stable across model providers and experiment settings.

These numbers highlight LLMs' limited ability to handle versioned data. The continual nature of ontology updates, combined with the fact that concept IDs may occur in training data with no indication of which version they are associated with, make it challenging to consistently eliminate old codes and incorporate current ones into these models.

5.6. Concept Depth

By taking a concept's depth in the SNOMED CT hierarchy as a proxy for its specificity, we find that LLMs are more likely to hallucinate codes for more specific concepts. As illustrated in Figure 3, we observe a highly statistically significant interaction between the depth of a ground truth concept label in the SNOMED CT hierarchy and the likelihood that a model will hallucinate its output for the corresponding span, though the effect size is small ($\rho = 0.09$, $p = 3 \times 10^{-22}$ averaged across all models). This likely reflects the fact that more general codes are better represented in the available training data, and therefore more likely to have been seen by models. The effect is strongest for Claude Sonnet ($\rho = 0.12$, $p = 1.5 \times 10^{-14}$); Sonnet is also the smallest model included in our experiments, so it is perhaps not surprising that it should exhibit the poorest recall for more specialized concepts.

For a discussion of concept depth as it relates to *non*-hallucinated model outputs, see Appendix C.

5.7. Invalid Codes

Without RAG, between 18% and 54% (Table 5) of the SCTIDs produced by LLMs are *invalid*; that is,

Model	Phrases with most hallucinations
GPT-5.2 (E2E)	WBC, PTT, Plt, Left chest tube†, ciprofloxacin†
GPT-5.2 (Mention)	WBC, RBC, Plt, MCH, MCV
GPT-5.2 (Oracle Span)	MCV, MCHC, MCH, RDW, BLOOD WBC
Claude Opus 4.6 (E2E)	RDW, MCH, AnGap, HCO3, PTT
Claude Opus 4.6 (Mention)	MCH, RDW, AnGap, TotBili, HCO3
Claude Opus 4.6 (Oracle Span)	MCHC, MCH, RDW, MCV, AnGap
Claude Opus 4.6 (RAG)	leptomeningeal enhancement†
Claude Sonnet 4.5 (E2E)	MCHC, MCV, MCH, RDW, AnGap
Claude Sonnet 4.5 (Mention)	Hct, RDW, MCHC, MCH, MCV
Claude Sonnet 4.5 (Oracle Span)	MCH, RDW, MCV, MCHC, AnGap
Claude Sonnet 4.5 (RAG)	foot drop

Table 4: Phrases which are associated with the greatest likelihood of a hallucinated concept ID. The 5 most-hallucinated phrases are shown for each model, separated by commas. †These phrases are never annotated in the ground truth data.

Model	Deprecated ↓
GPT-5.2 (E2E)	4.60%
GPT-5.2 (Mention)	4.60%
GPT-5.2 (Oracle Span)	4.34%
GPT-5.2 (RAG)	0.57%
Claude Opus 4.6 (E2E)	4.32%
Claude Opus 4.6 (Mention)	4.70%
Claude Opus 4.6 (Oracle Span)	4.50%
Claude Opus 4.6 (RAG)	0.63%
Claude Sonnet 4.5 (E2E)	3.61%
Claude Sonnet 4.5 (Mention)	4.31%
Claude Sonnet 4.5 (Oracle Span)	3.95%
Claude Sonnet 4.5 (RAG)	0.59%
Amazon Comprehend Medical	3.47%

Table 5: Macro-averaged proportion of outputs representing real but inactive SCTIDs.

Model	Invalid ↓
GPT-5.2 (E2E)	18.72%
GPT-5.2 (Mention)	27.71%
GPT-5.2 (Oracle Span)	36.88%
GPT-5.2 (RAG)	0.00%
Claude Opus 4.6 (E2E)	41.60%
Claude Opus 4.6 (Mention)	24.80%
Claude Opus 4.6 (Oracle Span)	34.07%
Claude Opus 4.6 (RAG)	0.02%
Claude Sonnet 4.5 (E2E)	54.46%
Claude Sonnet 4.5 (Mention)	34.07%
Claude Sonnet 4.5 (Oracle Span)	51.49%
Claude Sonnet 4.5 (RAG)	0.02%
Amazon Comprehend Medical	0.00%

Table 6: Macro-averaged proportion of outputs representing invalid SCTIDs.

beyond simply not existing in any current release of SNOMED CT, they could never exist as they violate one or more of the basic formatting requirements which SCTIDs are required to follow.

The overwhelming majority of these violations are due to invalid check digits. An SCTID must

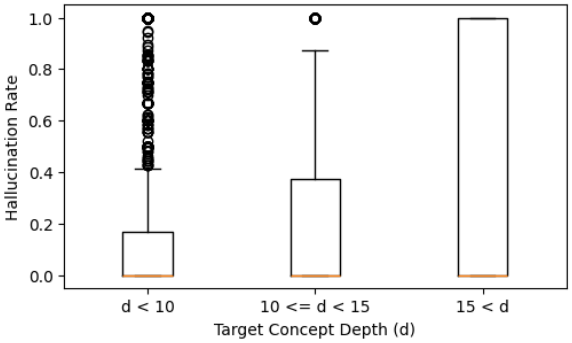


Figure 3: Hallucination rate vs target concept depth. Concepts which are deeper in the SNOMED CT hierarchy are associated with a higher risk of hallucination.

end with a single-digit checksum, which is incorrect in more than 99% of hallucinated codes in most settings. There are only two settings in which other error types make up a significant proportion of invalid codes. In the GPT-5.2 Oracle Span setting, roughly 4% of hallucinations include an invalid partition identifier, which indicates the type of component the SCTID refers to (concept, relationship, etc). In the GPT-5.2 Mention setting, 27% of hallucinations include an invalid partition identifier.

6. Related Work

Recent work has established robust frameworks for evaluating biomedical entity linking. Kartchner et al. (2023) benchmark nine models across dimensions including scalability and zero-shot robustness, while BioEL (Bathala et al., 2025) standardizes training across ontologies like UMLS (Bodenreider, 2004) and MeSH (Lipscomb, 2000). However, these works primarily target biomedicine and traditional discriminative architectures, leaving the

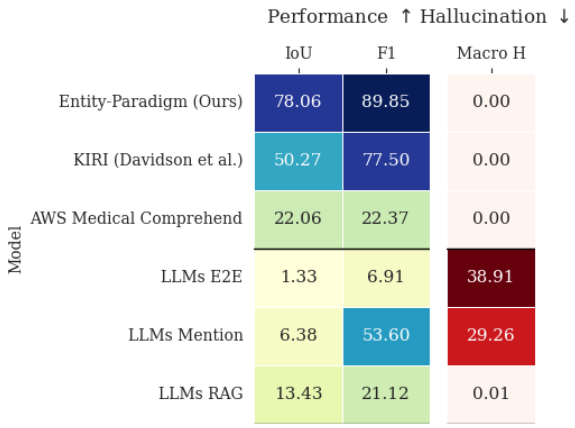


Figure 4: Summary of results grouped by setting, demonstrating high average hallucination rate and low average accuracy for generalist LLMs compared to task-specific models.

specific performance trade-offs and hallucination risks of LLMs in clinical settings largely unexplored.

Chang and Sung 2024 examine how SNOMED CT has been integrated into prior LLM research. They find that most research uses LLMs to produce contextual embedding vectors for entities, requiring fine-tuning on some downstream task. A majority of research integrates SNOMED CT in pretraining or fine-tuning, either lexically or in graph form.

Vishwanath et al. 2025 compare generalist frontier models against specialized clinical AI tools like OpenEvidence[8] and UpToDate Expert AI[9] on MedQA (Jin et al. 2020; USMLE-style questions) and HealthBench (Arora et al. 2025; clinical judgment). The results indicate that generalist LLMs outperform specialized tools on clinical judgment tasks, challenging marketing claims and emphasizing the need for transparent, independent evaluation before deployment of patient-facing workflows.

TrialMind, a pipeline built on GPT-4, was benchmarked against 100 published reviews involving over 2,220 studies (Wang et al., 2025). Medical experts preferred TrialMind's summaries over GPT-4, highlighting that the best clinical outcomes are achieved through a "hybrid" LLM-human collaboration (Wang et al., 2025).

Me-LLaMA, developed through continual pretraining and task-specific instruction tuning, demonstrated performance comparable to GPT-4 in diagnosing complex clinical cases (Xie et al., 2025). GPT-4 often outperform Google's Med-PaLM 2 (Singhal et al., 2025) in terms of both hallucination rate and human-centric measures like empathy and readability (Zeba et al., 2025).

Most studies investigating the performance of LLMs for autonomous medical coding (Soroush et al., 2024; Kwan, 2024; Hou et al., 2025; Yuan et al., 2025; Motzfeldt et al., 2025) focus on the task of clinical notes to medical billing codes using ICD dataset. While LLMs have been shown to generally struggle with autonomous medical coding (Soroush et al., 2024), Kwan 2024 propose a two-stage retrieve/rerank framework to improve performance. However, the restrictions on the input format render it incompatible with real world clinical notes. Yuan et al. 2025 propose adaptations to enhance the performance of LLMs and fix errors from hierarchical misalignments. In contrast, the focus of this paper is SNOMED coding. While ICD-9 and ICD-10 coding involve document level information and consolidation, SNOMED coding is more granular and ties medical concepts directly to spans of text.

7. Conclusion

We have demonstrated that medical entity linking against a controlled ontology provides good testbed for distinguishing hallucinations from other kinds of erroneous LLM output. We show that at least 20%, and in some cases nearly 55% of the SNOMED IDs (SCTIDs) produced by frontier LLMs on an entity linking task do not exist in the ontology these models were instructed to use. Of these erroneous codes, roughly one-third are fabrications which do not satisfy the formatting and other validation requirements to be interpreted as valid SCTIDs and are therefore highly unlikely to have been seen during training. Roughly 4.5% of erroneous codes *were* valid at some point in time, but have since been deprecated, highlighting that these models have difficulty keeping track of differences between ontology versions and are prone to conflating them. All models demonstrate a similar tendency to conflate the core SNOMED CT ontology with SNOMED CT *extensions* produced by national or corporate healthcare organizations. Since these extension codes *are* technically valid SCTIDs, they represent a particularly subtle form of error that may be difficult to catch if one is not specifically looking for it. Overall, these results demonstrate that LLMs still lack the reliability required for high-stakes medical applications, and highlight areas of particular concern for prospective users of LLM-based medical coding systems. As a future direction, we believe Model Context Protocol (MCP) tools[10] which are specialized for entity linking offer a better way to align LLM outputs with medical knowledge (El-Sayed et al., 2025) and to assist with medical coding tasks.

[8] https://www.openevidence.com/

[9] https://www.wolterskluwer.com/en/solutions/uptodate

[10] https://www.anthropic.com/news/model-context-protocol

8. 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, Johannes Heidecke, and Karan Singhal. 2025. Healthbench: Evaluating large language models towards improved human health.

Prasanth Bathala, Christophe Ye, Batuhan Nursal, Shubham Lohiya, David Kartchner, and Cassie S. Mitchell. 2025. BioEL: A comprehensive python package for biomedical entity linking. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 1709–1721, Albuquerque, New Mexico. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research, pages 267–270.

Eunsuk Chang and Sumi Sung. 2024. Use of snomed ct in large language models: Scoping review. JMIR Med Inform, 12:e62924.

Rory Davidson, Will Hardman, Guy Amit, Yonatan Bilu, Vincenzo Della Mea, Aleksandr Galaida, Irena Girshovitz, Mikhail Kulyabin, Mihai Horia Popescu, Kevin Roitero, Gleb Sokolov, and Chen Yanover. 2025. Snomed ct entity linking challenge. Journal of the American Medical Informatics Association, 32(9):1397–1406.

Zag ElSayed, Craig Erickson, and Ernest Pedapati. 2025. Mcp-ai: Protocol-driven intelligence framework for autonomous reasoning in healthcare.

Will Hardman, Mark Banks, Rory Davidson, Donna Truran, Nindya Widita Ayuningtyas, Hoa Ngo, Alistair Johnson, and Tom Pollard. 2025. SNOMED CT Entity Linking Challenge. PhysioNet. Version 1.1.0.

Zhen Hou, Hao Liu, Jiang Bian, Xing He, and Yan Zhuang. 2025. Enhancing medical coding efficiency through domain-specific fine-tuned large language models. npj Health Systems, 2(1):14.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst., 43(2).

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. arXiv preprint arXiv:2009.13081.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. MIMIC-IV-Note: Deidentified free-text clinical notes. PhysioNet. Version 2.2.

David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie Mitchell. 2023. A comprehensive evaluation of biomedical entity linking models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 14462–14478, Singapore. Association for Computational Linguistics.

Keith Kwan. 2024. Large language models are good medical coders, if provided with tools.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). Bulletin of the Medical Library Association.

Subhankar Maity and Manob Jyoti Saikia. 2025. Large language models in healthcare and medical applications: A review. Bioengineering (Basel), 12(6):631.

Andreas Geert Motzfeldt, Joakim Edin, Casper L. Christensen, Christian Hardmeier, Lars Maaløe, and Anna Rogers. 2025. Code like humans: A multi-agent solution for medical coding. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22612–22627, Suzhou, China. Association for Computational Linguistics.

Wenbo Pan, Jie Xu, Qiguang Chen, Junhao Dong, Libo Qin, Xinfeng Li, Haining Yu, and Xiaohua Jia. 2025. Can llms refuse questions they do not know? measuring knowledge-aware refusal in factual tasks.

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.

Ali Soroush, Benjamin S Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. 2024. Large language models are poor medical coders — benchmarking of medical code querying. NEJM AI.

Krithik Vishwanath, Mrigayu Ghosh, Anton Alyakin, Daniel Alexander Alber, Yindalon Aphinyanaphongs, and Eric Karl Oermann. 2025. Generalist large language models outperform clinical tools on medical benchmarks.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. 2025. Accelerating clinical evidence synthesis with large language models. npj Digital Medicine, 8(1):509.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. 2025. Medical foundation large language models for comprehensive text analysis and beyond. NPJ digital medicine, 8(1):141.

Moy Yuan, Han-Chin Shing, Mitch Strong, and Chaitanya Shivade. 2025. Toward reliable clinical coding with language models: Verification and lightweight adaptation. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 173–184, Suzhou (China). Association for Computational Linguistics.

Musarrat Zeba, Abdullah Al Mamun, Kishoar Jahan Tithee, Debopom Sutradhar, Mohaimenul Azam Khan Raiaan, Saddam Mukta, Reem E. Mohamed, Md Rafiqul Islam, Yakub Sebastian, Mukhtar Hussain, and Sami Azam. 2025. Mitigating hallucinations in healthcare llms with granular fact-checking and domain-specific adaptation.

A. Reannotation methodology

As noted in Section 3.1, the original SNOMED CT entity linking challenge (Davidson et al., 2025) suffers from inconsistent labelling, such that two identical phrases in different texts may be annotated with a concept link in one case but left unannotated in the other. This provides an inconsistent training signal for models which require training data, and imposes an upper limit on the accuracy a model can achieve on the test set, as it comes down to random chance whether a given semantically-correct concept link will have been included among the ground truth spans. For this reason, the upper limit of scores achievable on the original test set is likely much lower than the theoretical limit of 1.0 IoU, since past a certain point, improving a model's score would require knowledge of which spans were actually covered by the ground truth annotations so that the model could avoid producing semantically-correct outputs which happen to have been missed by the annotators.[11]

Furthermore, the challenge utilized a simplified subset of SNOMED CT (May 2023 International Edition) focusing only on three sub-hierarchies (findings, procedures, and body structures), which limits the breadth of model behaviours it can be used to evaluate.

These factors motivated our decision to reannotate the dataset for this work. We use an automated system to assign concept links, which ensures a greater degree of consistency across documents and eliminates the possibility of inter-annotator disagreement. The system used to assign concept links is trained on a corpus of clinical notes which have been manually annotated by clinicians: its training inputs comprise 1,137,651 manual annotations covering 380,588 SCTIDs. We employ a broader set of semantic types than the original dataset in order to capture a more comprehensive range of clinical entities. The semantic types used in our evaluation were: disorder, procedure, finding, morphologic abnormality, body structure, observable entity, and regime/therapy. This expanded scope allows for a more thorough assessment of models' ability to handle diverse medical terminology across the full clinical spectrum, including pathological states, therapeutic interventions, clinical observations, anatomical components, and treatment regimens. Medicinal products and substances (e.g., ciprofloxacin) are excluded from this evaluation, as they fall outside the semantic types used.

Figure 5 shows a sample of sentences where the set of annotations has changed as a result of our reannotation.

Since concept links were assigned automatically, we add an LLM filtering step to flag and remove links that may have been assigned in error. In total, Claude Haiku 4.5 flags 13% of the automatically applied concept links for removal. We present a random sample of the removed links in Table 7. We note that while the LLM filtering does remove some legitimate mistakes (e.g. "he admitted" labelled as a hospital admission), it can also be overzealous about removing valid links when they occur in a negated context ("no vomiting") or when the concept wording does not match its expectations (insisting that "VII" is a cranial nerve, not a facial nerve, when in fact it is both).

B. Embedding-Weighted IoU

Standard entity linking metrics such as IoU and F1 conflate a model's ability to produce semantically-valid concept links with its ability to follow a specific annotation style. A model which annotates `left and right` as the concept `Right and left (qualifier value)` is as correct, in semantic terms, as one which labels `left` as `Left (qualifier value)` and `right` as `Right (qualifier value)`. However, in terms of IoU or F1, segmenting the phrase in this way incurs the same penalty as if the model had mislabelled it entirely.

We propose an embedding-weighted IoU score (EWIoU) which awards partial points based on the cosine similarity between dense representations of the observed and expected concepts as a means of decoupling semantic correctness from segmentation or specificity.

If P_c is the set of characters which were predicted to belong to some concept c, and G_c is the set of characters labelled with concept c in the ground truth data, IoU computes the following quantity:

$$\mathrm{IoU}_c = \frac{|P_c \cap G_c|}{|P_c \cup G_c|}$$

Now let $\mathbf{e}_c$ be a dense embedding representing concept c, and likewise let $\mathbf{e}_{\mathrm{char}}$ be the embedding for the concept assigned to the character char by some model. (We use the same OpenAI `text-embedding-3-large` embeddings as in the RAG retriever.) Then the embedding-weighted IoU is:

$$\mathrm{EWIoU}_c = \frac{\sum_{\mathrm{char} \in P_c \cup G_c} \mathbf{e}_c \cdot \mathbf{e}_{\mathrm{char}}}{|P_c \cup G_c|}$$

[11] It is also for this reason that we report the score of our `entity-paradigm` system on the test set, even though it is the model used to prepare the test set and is therefore unfairly advantaged. Our score represents an approximate upper bound on the scores that should be achievable on this dataset when a model's level of granularity, phrase segmentation choices, set of known concepts, etc are perfectly matched to the test set.

Original: There exists some crowded vasculature in the
left lower lobe area in retrocardiac position possibly suggesting some atelectasis but acute discrete
41224006 | Structure of lower lobe of left lung (body structure) | 46621007 | Atelectasis (disorder) |
parenchymal infiltrates identifying a pneumonia cannot be seen.
231287002 | Infiltration (procedure) | 233604007 | Pneumonia (disorder) |

Updated: There exists some crowded vasculature in the
59820001 | Blood vessel structure (body structure) |
left lower lobe area in retrocardiac position possibly suggesting some atelectasis but acute discrete
41224006 | Structure of lower lobe of left lung (body structure) | 46621007 | Atelectasis (disorder) |
parenchymal infiltrates identifying a pneumonia cannot be seen.
409609008 | Radiologic infiltrate of lung (disorder) | 233604007 | Pneumonia (disorder) |

Original: BLOOD Glucose -219*
33747003 | Glucose measurement, blood (procedure) |
UreaN -19
105011006 | Blood urea nitrogen measurement (procedure) |
Creat -0.9
70901006 | Creatinine measurement (procedure) |
Na -138
312469006 | Blood sodium measurement (procedure) |
K-4.3
Cl -104
104589004 | Chloride measurement, blood (procedure) |
HCO3 -26
312471006 | Blood bicarbonate measurement (procedure) |
AnGap -12
25469001 | Anion gap measurement (procedure) |

Updated: BLOOD Glucose -219*
33747003 | Glucose measurement, blood (procedure) |
UreaN -19
105011006 | Blood urea nitrogen measurement (procedure) |
Creat -0.9
113075003 | Creatinine measurement, serum (procedure) |
Na -138
104934005 | Sodium measurement, serum (procedure) |
K -4.3
271236005 | Serum potassium measurement (procedure) |
Cl- 104
271238006 | Serum chloride measurement (procedure) |
HCO3- 26
312471006 | Blood bicarbonate measurement (procedure) |
AnGap -12
271057005 | Serum anion gap measurement (procedure) |

Figure 5: Sample of differences resulting from reannotation. Our update provides broader coverage by adding new links for "vasculature" and "K". We replace an incorrect mapping for "infiltrates" with a more correct, and longer, mapping for "parenchymal infiltrates". Our concept assignments are more internally consistent: blood measurements are labelled as such in our output, where the original occasionally uses a more generic measurement concept.

Table 8 compares the plain and embedding-weighted IoU scores for all models in all settings. Note that the EWIoU score for `entity-paradigm` is identical to its IoU score. This is expected, as it is the model that was used to update the data set with more comprehensive concept links, and it should trivially reproduce its own choices regarding phrase segmentation and level of detail. Thus it does not benefit from a more relaxed evaluation.

Other models see significant improvements, however: among the task-specific models, KIRI gains +13 percentage points and Amazon Comprehend

Concept ID	FSN	Context	Justification
119265000	Assisting (procedure)	Her daughter [...] was able to **assist** with much of the history	Assist means support/help, not a medical procedure. Context is supportive role, not clinical intervention.
56052001	Facial nerve structure (body structure)	V, **VII**: Facial strength and sensation intact and symmetric	VII is cranial nerve designation, not facial nerve body structure.
32485007	Hospital admission (procedure)	He **admitted** that he had lost 10 lbs	"admitted" used as past tense verb describing weight loss acknowledgment, not hospital admission procedure.
420227002	Recommendation to (procedure)	per palliative **recommendations** will be discharged home	Concept too specific; "recommendations" refers to discharge plans, not procedure recommendations.
58000006	Patient discharge (procedure)	**DISCHARGE** LABS:	DISCHARGE in section heading refers to test type, not patient discharge procedure.
74262004	Oral cavity structure (body structure)	Lansoprazole **Oral** Disintegrating Tab 15 mg PO DAILY	Route of administration mismatched with anatomical structure concept.
31156008	Structure of left half of body (body structure)	you had pain on your **left** back and under your ribs	Concept 'left' is overly broad; matching text specifies 'left back' only.
307818003	Weight monitoring (regime/therapy)	Remember to **weigh yourself every morning**	FSN is literal instruction, not therapeutic procedure concept. Misaligned with clinical intent.
422400008	Vomiting (disorder)	Her headaches ia also associated with nausea and dry heaves, no **vomiting**.	Explicit negation: "no vomiting" contradicts vomiting concept despite guideline allowing negation.

Table 7: Random sample of annotations which were removed by LLM filtering, and justification provided by LLM for removal.

Task-Specific Models		
Model	**IoU** ↑	**EWIoU** ↑
entity-paradigm (Ours)	78.06%	78.06%
KIRI	50.49%	63.93%
Amazon Comprehend Medical	22.06%	41.68%
General-Purpose Frontier LLMs		
Model	**IoU** ↑	**EWIoU** ↑
GPT-5.2 (E2E)	0.15%	2.93%
Claude Opus 4.6 (E2E)	2.23%	12.66%
Claude Sonnet 4.5 (E2E)	1.60%	10.52%
GPT-5.2 (Mention)	5.39%	25.47%
Claude Opus 4.6 (Mention)	8.98%	27.34%
Claude Sonnet 4.5 (Mention)	4.78%	17.05%
GPT 5.2 (RAG)	11.66%	41.14%
Claude Opus 4.6 (RAG)	**14.56%**	**45.39%**
Claude Sonnet 4.5 (RAG)	14.08%	43.12%
Oracle Setting		
Model	**IoU** ↑	**EWIoU** ↑
entity-paradigm (Ours; Oracle Span)	84.45%	84.45%
GPT-5.2 (Oracle Span)	6.69%	38.83%
Claude Opus 4.6 (Oracle Span)	**12.36%**	**40.43%**
Claude Sonnet 4.5 (Oracle Span)	8.02%	29.67%

Table 8: Embedding-Weighted IoU scores for all models.

Medical gains +19. For the generalist LLMs, the largest absolute gains are seen in the RAG setting, where all models gain upwards of +30 points; the absolute gain is more modest in the other settings, but is still quite large in relative terms, representing nearly a twenty-fold improvement over regular IoU in the worst-performing setting.

This discrepancy highlights that all models produce a relatively large number of outputs which are semantically close to the target concept even when they do not match the target exactly. This demonstrates a shortcoming of exact-match metrics like IoU in settings where multiple annotations are potentially valid, and where correctness may depend more on specific annotation conventions than on the underlying semantics of the concept link. We

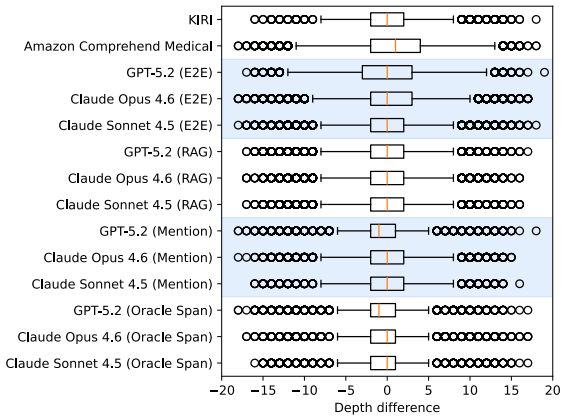


Figure 6: Difference between depth of predicted and target concepts in the SNOMED CT hierarchy.

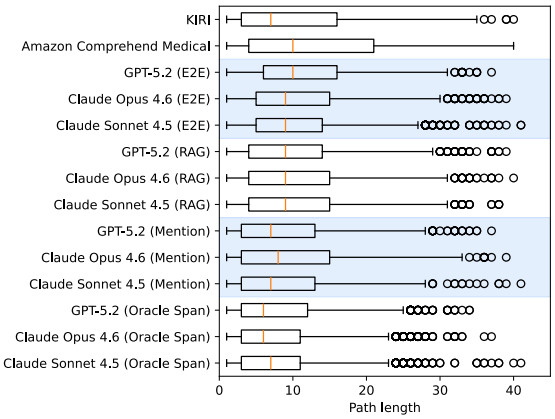


Figure 7: Length of shortest path from predicted to target concept in the SNOMED CT hierarchy.

suggest that fuzzy evaluations like EWIoU may give a more accurate view of semantic accuracy in such settings.

C. Additional Remarks on Concept Depth

When a model produces a real, but incorrect, concept, it is roughly equally likely to be deeper in the hierarchy (i.e. more specific) than the target as it is to be shallower (i.e. more generic; Figure 6). Across all models, the median difference in depth between observed and expected concepts is zero, with the exception of Amazon Comprehend Medical, where the predicted concept is one level deeper than the target on average, and GPT-5.2 (Mention and Oracle Span) where it is one level shallower.

To further investigate the nature of non-hallucinated errors, we compute the length of the path linking an erroneous prediction to the corresponding target in the SNOMED CT hierarchy. We walk up the hierarchy from the predicted concept to the lowest common ancestor of the target and the prediction, then back down to the target. A boxplot of the resulting path lengths is shown in Figure 7.

The shortest mean path lengths are found in the Oracle Span setting. This is unsurprising, as knowledge of the target span also provides information about the expected level of granularity of the target concept. The longest path lengths are incurred by Amazon Comprehend Medical. Where the other models will label ambiguous mentions (e.g. “disease”) with accurate but generic concepts (`Disease (disorder)`), the Amazon model attempts to resolve these to the specific underlying disorder based on context clues. Ignoring the accuracy of this disambiguation, which is outside the scope of the present work, this approach produces concepts at a much different level of granularity than any of the other approaches, which is reflected by the longer path lengths.

D. Prompt Engineering and Optimization Protocol

To ensure a robust comparison between generalist LLMs and the task-specific commercial benchmark, we employed a systematic optimization process for all prompt-based models. Rather than relying on a single static prompt, we treated prompt design as a critical experimental variable:

1. **Iterative Optimization**: We conducted an iterative search for the optimal prompt structure using a small, held-out validation set. This included testing various instruction formats, persona assignments (e.g, "You are a medical terminology expert .."), and the inclusion of explicit logical constraints.

2. **Reasoning Strategies**: To ensure the generalist models were evaluated at their best capability, we tested several reasoning configurations, including chain-of-thought (CoT). Our optimization process showed that for this task, the inclusion of intermediate reasoning steps or extended "thinking" modes introduced unnecessary variance and lower overall accuracy. Therefore, we utilized optimized direct-response prompts, which provided the most robust and competitive results for the LLMs.

Our `entity-paradigm` model, the commercial offering by *emtelligent*, used in this study is a proprietary, task-specific architecture that does not rely on prompting. It is included as a ’Performance Oracle’ to provide a high-water mark for the task. By ensuring that the generalist LLM prompts were fully optimized, we can more accurately quantify

the remaining performance gap between general-purpose models and specialized, state-of-the-art solutions.

Oracle Span Prompt Template

```
You are a medical terminology expert specializing in SNOMED CT (
    Systematized Nomenclature of Medicine - Clinical Terms).

Given a clinical note and a list of medical entity mentions extracted from
     it, your task is to provide the SNOMED CT concept ID for each entity
    mention.

<clinical_note>
{{ note_text }}
</clinical_note>

## Entity Mentions to Link
{% for entity in entities %}
- ID {{ loop.index0 }}: "{{ entity.mention_text }}" (position {{ entity.
    start }}-{{ entity.end }})
{% endfor %}

## Instructions
For each entity mention above, provide the most appropriate SNOMED CT
    concept ID.

Reason about the correct SNOMED CT concept for each entity  in <scratchpad
    > tags BEFORE you start writing JSON. Your <scratchpad> must be short
    and concise, in very few words. Once you begin the JSON array, do NOT
    interrupt it with any commentary, reasoning, or corrections - output
    the complete array in one unbroken block inside the <json> tags.

STRICT FORMAT RULES:
1. The response MUST contain exactly one JSON array inside the <json> tags
2. The array MUST contain exactly {{ entities|length }} objects, one per
    entity
3. Every object MUST have exactly two keys: "entity\_id" (integer) and "
    concept\_id" (string)
4. Do NOT omit "entity_id" from any object. Every single object MUST
    include both "entity_id" AND "concept_id"
5. The "entity_id" values MUST be {{ range(entities|length)|list|join(',
    ') }} in that exact order
6. The "concept_id" MUST be a numeric SNOMED CT code as a string (e.g.
    "89187006")

Example for 3 entities:
<json>
[{"entity_id": 0, "concept_id": "89187006"}, {"entity_id": 1, "concept_id
    ": "38341003"}, {"entity_id": 2, "concept_id": "73211009"}]
</json>

Now output the JSON array for all {{ entities|length }} entities.
```

Figure 8: Prompt template for the **Oracle Span** setting.

End-to-End Prompt Template

```
You are a medical terminology expert specializing in SNOMED CT (
    Systematized Nomenclature of Medicine - Clinical Terms).

Given a clinical note, your task is to:
1. Identify all medical entities (anatomical structures, clinical findings
    , procedures, substances, disorders, etc.)
2. For each entity, provide its exact character positions in the text and
    the corresponding SNOMED CT concept ID.

<clinical_note>
{{ note_text }}
</clinical_note>

## Instructions
Extract every medical entity from the clinical note and link each to its
    SNOMED CT concept ID.

Reason about the correct mention and it's SNOMED CT concept inside <
    scratchpad> tags BEFORE you start writing JSON. Your <scratchpad> must
     be short and concise, in very few words. Once you begin the JSON
    array, do NOT interrupt it with any commentary, reasoning, or
    corrections - output the complete array in one unbroken block inside
    the <json> tags.

Respond with ONLY a JSON array, no other text. Each element must have:
- "text": the exact entity mention as it appears in the note - must be
    present
- "start": INT character offset where the mention begins (0-indexed from
    the start of the note text) - must be present
- "end": INT character offset where the mention ends (exclusive) - must be
     present
- "concept_id": the SNOMED CT concept ID as a string - this field is
    mandatory for every entry. If you are uncertain or a concept id is not
     found, provide 000000000 as concept ID. Never omit this field.

Example format:
<json>
[{"text": "airway", "start": 1078, "end": 1084, "concept_id": "89187006"},
     {"text": "Penicillins", "start": 1168, "end": 1179, "concept_id":
    "764146007"}]
</json>

Be thorough and ensure every object in the array contains all four keys.
    Include all medical terms: anatomical sites, diagnoses, symptoms,
    procedures, medications, lab values, and clinical observations. "start
    " and "end" fields must be integers to denote character offsets.
```

Figure 9: Prompt template for the **End-to-End** setting.

Mention-based linking Prompt Template

```
You are a medical terminology expert specializing in SNOMED CT (
    Systematized Nomenclature of Medicine - Clinical Terms).

Given a clinical note, your task is to:
1. Identify all medical entities (anatomical structures, clinical findings
    , procedures, substances, disorders, etc.)
2. For each entity, provide the exact mention text as it appears in the
    note, the SNOMED CT concept ID, and which occurrence of that text in
    the note it refers to.

<clinical_note>
{{ note_text }}
</clinical_note>

## Instructions
Extract every medical entity from the clinical note and link each to its
    SNOMED CT concept ID.

Reason about the correct mention and it's SNOMED CT concept inside <
    scratchpad> tags BEFORE you start writing JSON. Your <scratchpad> must
     be short and concise, in very few words. Once you begin the JSON
    array, do NOT interrupt it with any commentary, reasoning, or
    corrections - output the complete array in one unbroken block inside
    the <json> tags.

Respond with ONLY a JSON array, no other text. Each element must have:
- "text": the exact entity mention as it appears in the note (preserve
    original casing)
- "concept_id": the SNOMED CT concept ID as a string
- "occurrence": which occurrence of this exact text in the note (1 = first
    , 2 = second, etc.). Count occurrences case-insensitively.

For example, if "chest" appears 4 times in the note and the 3rd occurrence
     refers to SNOMED CT concept 51185008:
<json>
[{"text": "chest", "concept_id": "51185008", "occurrence": 3}]
</json>

Be thorough. Include all medical terms: anatomical sites, diagnoses,
    symptoms, procedures, medications, lab values, and clinical
    observations. If the same mention appears multiple times and each
    should be linked, include a separate entry for each occurrence.
```

Figure 10: Prompt template for the **Mention** setting.

Mention Extraction Prompt Template for RAG

```
You are a medical terminology expert specializing in SNOMED CT (
    Systematized Nomenclature of Medicine - Clinical Terms).

Given a clinical note, your task is to identify all medical entities (
    anatomical structures, clinical findings, procedures, substances,
    disorders, etc.) and return each as a mention with its occurrence
    number.

<clinical_note>
{{ note_text }}
</clinical_note>

## Instructions
Extract every medical entity from the clinical note. Do NOT attempt to
    predict SNOMED CT concept IDs - only extract the mention text and
    occurrence number.

Reason about the correct mentions inside <scratchpad> tags BEFORE you
    start writing JSON. Your <scratchpad> must be short and concise, in
    very few words. Once you begin the JSON array, do NOT interrupt it
    with any commentary, reasoning, or corrections - output the complete
    array in one unbroken block inside the <json> tags.

Respond with ONLY a JSON array, no other text. Each element must have:
- "text": the exact entity mention as it appears in the note (preserve
    original casing)
- "occurrence": which occurrence of this exact text in the note (1 = first
    , 2 = second, etc.). Count occurrences case-insensitively.

For example, if "chest" appears 4 times in the note and the 3rd occurrence
     is a medical entity:
<json>
[{"text": "chest", "occurrence": 3}]
</json>

Be thorough. Include all medical terms: anatomical sites, diagnoses,
    symptoms, procedures, medications, lab values, and clinical
    observations. If the same mention appears multiple times and each
    should be linked, include a separate entry for each occurrence.
```

Figure 11: Prompt template for the **Mention extraction** (RAG Pt. 1).

Retrived-concept Linking Prompt Template for RAG

```
You are a medical terminology expert specializing in SNOMED CT (
    Systematized Nomenclature of Medicine - Clinical Terms).

Given a clinical note and a list of medical entity mentions (each with
    candidate SNOMED CT concepts), your task is to select the best
    matching SNOMED CT concept ID for each mention from its candidate list.


<clinical_note>
{{ note_text }}
</clinical_note>

## Mentions and Candidates
{% for mention in mentions %}
### Mention {{ mention.id }}: "{{ mention.text }}"
Candidates:
{% for c in mention.candidates %}
- {{ c.snomed_id }}: {{ c.fsn }}
{% endfor %}
{% endfor %}

## Instructions
For each mention, select the single best-matching SNOMED CT concept from
    its candidate list. Consider the clinical context in the note to
    disambiguate.

Reason about each mention inside <scratchpad> tags BEFORE you start
    writing JSON. Your <scratchpad> must be short and concise. Once you
    begin the JSON array, do NOT interrupt it with any commentary - output
     the complete array in one unbroken block inside the <json> tags.

Respond with a JSON array with exactly {{ mentions|length }} elements, one
     per mention. Each element must have:
- "mention_id": the mention number (integer)
- "concept_id": the selected SNOMED CT concept ID as a string

If none of the candidates are appropriate, use "000000000" as the
    concept_id.

<json>
[{"mention_id": 0, "concept_id": "89187006"}, ...]
</json>
```

Figure 12: Prompt template for the **Concept selection** (RAG Pt. 2).

Author Index

AAlAbdulsalam, Abdulrahman Khalifa, 125
Afzal, Muhammad Umar, 350
Akumu, Tanya, 262
Alex, Beatrice, 92
Alonso Alemany, Laura, 270
AlZaabi, Adhari, 125
Andaur Navarro, Constanza L., 174
Aramaki, Eiji, 222, 324
Aryoyudanta, Bayu, 163

Bader, Cari, 19
Bandhey, Harsh, 27
Beaucoté, Victor, 57
Beliveau, Vincent, 313
Ben Abacha, Asma, 1, 19
Bethard, Steven, 336
Bhagwanani, Satyam, 381
Bodenstab, Nathan, 19
Boerma, Tessel, 103
Bonnevie, Rasmus, 313
Born, Logan, 394
Boylan, Jack, 290
Bratières, Sébastien, 113
Brunello, Florencia Luciana, 270

Calixto, Iacer, 113
Carbonari, Valentina, 235
Casey, Arlene, 92
Cassidy, Lauren, 290
Charron, Cyril, 57
Corbeil, Jean-Philippe, 19
Cui, Haochen, 141
Cunha, Luís Filipe, 300

D'Ambrosi, Leonardo, 369
Deepak, Pasumarthy, 228
Diaz Galiano, Manuel Carlos, 73
Dineen, Jacob, 350
Dligach, Dmitriy, 212
Dobson, Richard J., 381
Donelly, Hayoung K., 27
Dukmak, Marya, 174

E, Liyuan, 12
Eisenberg, Matthew A., 336
El Fatimi, Meryem, 49
Elliott, Desmond, 313
Eslami, Behnaz, 212
Eyre, David W., 141

Feng, Kun, 12
Fernandes, Ana Luisa, 300
Fraccaro, Marco, 313

Gangavarapu, Haritha, 257
Ganz, Melanie, 313
Gérardin, Christel, 57
Geva, Alon, 336
Gholipour Ghalandari, Demian, 290
Grover, Claire, 92
Guzzi, Pietro Hiram, 235

Hanna, Ameen Abu, 113
Hao, Ziqi, 192
Hokamp, Chris, 290
Horiguchi, Koki, 222
Hwang, Kyomin, 66
Hwang, Samuel, 141
Hwang, Sy, 27

Ingala, Silvia, 313

JeebAllah, Riham, 125
Joseph, Adrien, 57

Kadusabe, Provia, 290
Kajiwara, Tomoyuki, 222
Kambhatla, Nishant, 394
Karim, A H M Rezaul, 153
Kaushik, Abhishek, 290
Kelly, Daniel P., 336
Khong, Benjamin, 141
Kim, Sue Hyon, 27
King, Andrew J., 27
Kocbek, Primoz, 235
Kondadadi, Rishik, 38
Krishnamurthy, Parameswari, 228, 251
Kubasova, Uliyana, 394
Kwak, Nojun, 66

Lawless, Fiona, 290

Lazrak, Noussair, 262
Lee, Yoonjae, 27
Leeuwenberg, Artuur, 174
Legrand, Paul, 381
Lekadir, Karim, 262
Li, Jielin, 12
Li, Yingya, 336
Li, Zhao, 12
Liu, Pengbo, 192
Luk, Augustine, 141

Madden, Kate, 336
Marimuthu, Sushvin, 228
Martín-Valdivia, María-Teresa, 73
Michalopoulos, George, 19
Miller, Timothy A., 336
Miranda, Michele, 113
Mishra, Nishant, 113
Molino-Piñar, Lucas, 73
Mowery, Danielle L., 27
Munjal, Prateek, 83
Murphy, Rachel, 113

Naguib, Marco, 57
Nainia, Ayoub, 49, 262
Névéol, Aurélie, 57
Nielsen, Michael Bachmann, 313
Nishiyama, Tomohiro, 324
Noor, Kawsar, 381

O'Connell, Timothy W., 394
O'Connor, Karen, 27
Odijk, Jan, 103
Ortega, John E., 38

Pehrson, Lea Marie, 313
Peng, Xinzhe, 12
Pitcher, Katherine S., 27
Pradhan, Alyssa, 141
Pusapati, Siva Satyanarayana Raju, 183

Rachman, Mikie, 163
Riaz, Irbaz, 350

Saban, Michael T., 212
Sahoo, Akshitaa, 244
Sarfraz, Muhammad Uzair, 350
Sarkar, Anoop, 394
Savova, Guergana, 336
Schiavone, Alice, 313
Schraagen, Marijn, 103
Sekizaki, Ryo, 281
Setiawan, I Made Agus, 163
Shao, Zhihao, 281
Sharma, Dipti Misra, 136
Shimizu, Seiji, 324
Siahbani, Maryam, 394
Silvano, Purificação, 300
Singh, Ankit, 183
Srinivasan, Adarsh, 350
Sriram, Swetha Krishna, 244
Stiglic, Gregor, 235
Sugiyama, Seiji, 222

Tang, Yuxuan, 12
Tannier, Xavier, 57
Tashiro, Yuki, 324
Teruel, Milagro, 270
Thomas, Jerrin John, 251
Tobin, Richard, 92
Tootooni, Samie, 212

Urbanowicz, Ryan J., 27
Uzuner, Özlem, 153

Vacariu, Andrei, 394
Vaish, Aashwin, 136
Valiev, Airat A., 200
van Boheemen, Jelte, 103
Veltri, Pierangelo, 235
Vertegaal, Xander, 103
Villata, Serena, 270

Wakamiya, Shoko, 222, 324
Whalley, Heather, 92
Whiteley, William, 92

Xu, Justin, 141

Yaghoubi, Arsalan, 212
Yamasaki, Toshihiko, 281
Yan, Xinlan, 113
Yi, Shengzhou, 281
Yim, Wen-wai, 1
Yuan, Kevin, 141
Yuliana, Maria, 163

Zahir, Jihad, 49
Zhang, Zizheng, 141
Zhou, Ben, 350
Ziletti, Angelo, 369