

NLP METHODS FOR EMBEDDING REAL-WORLD CLINICAL KNOWLEDGE IN LLMS: TOWARDS RESPONSIBLE, TRANSFORMATIVE, MEDICAL AI

Graciela Gonzalez-Hernandez, PhD

*Department of Computational Biomedicine, Cedars-Sinai Medical Center,
Los Angeles, CA, USA*

Email: Graciela.GonzalezHernandez@csmc.edu

Carl Berdahl, MD

*Cedars-Sinai Medical Center,
Los Angeles, CA, USA*

Email: Carl.Berdahl@csmc.edu

Brian Chapman, PhD

*UT Southwestern Medical Center,
Dallas, TX, USA*

Email: brian.chapman@utsouthwestern.edu

Wendy W. Chapman, PhD

*UT Southwestern Medical Center,
Dallas, TX, USA*

Email: wendy.chapman@utsouthwestern.edu

Sumeet Chugh, MD

*Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System,
Los Angeles, CA, USA*

Email: Sumeet.Chugh@csmc.edu

Ian M. Campbell, MD, PhD

*Children's Hospital of Philadelphia,
Philadelphia, PA, USA*

Email: campbellim@chop.edu

Timothy Daskivich, MD

*Cedars-Sinai Medical Center,
Los Angeles, CA, USA*

Email: Timothy.Daskivich@csmc.edu

Hongfang Liu, PhD

*Dell Medical School, UT Austin,
Austin, TX, USA*

Email: hongfang.liu@austin.utexas.edu

Abeed Sarker, PhD

Emory University,

Atlanta, GA, USA

Email: abeed@dbmi.emory.edu

Davy Weissenbacher, PhD

Department of Computational Biomedicine, Cedars-Sinai Medical Center,

Los Angeles, CA, USA

Email: Davy.Weissenbacher@cshs.org

This session examines natural language processing methods for embedding, retrieving, structuring, evaluating, and governing real-world clinical knowledge in large language models. The seven accepted papers address clinical surveillance, patient-reported outcomes, longitudinal EHR reconstruction, synthetic clinical trajectories, medical search, evidence synthesis, and diagnostic reasoning. Across these settings, the papers move beyond final-answer accuracy toward explicit and auditable representations of timelines, patient states, reasoning traces, search behavior, and observability. They also highlight the complementarity of narrative notes, structured records, patient-generated information, case reports, and biomedical literature. The session identifies shared opportunities for cross-institutional benchmarks, provenance-aware evaluation, patient-centered knowledge integration, prospective human-AI assessment, and governance of updating and escalation.

Keywords: clinical NLP; large language models; medical knowledge; retrieval; clinical reasoning; evaluation; patient-centered data; responsible AI

1. Introduction

Large language models (LLMs) are shifting biomedical natural language processing from a field centered primarily on extraction, classification, and concept normalization toward one increasingly concerned with generation, reasoning, evidence retrieval, and interactive decision support. In medicine, realizing this potential requires methods that embed real-world clinical knowledge in LLM-based systems in ways that are explicit, evidence-grounded, current, and appropriate to the clinical context. Real-world clinical knowledge is distributed across guidelines, ontologies, electronic health record (EHR) narratives, structured observations, registries, biomedical literature, patient-reported outcomes, and patient communications.

These sources provide complementary but partial representations of clinical phenomena, patient trajectories, and care processes. They differ in authority, provenance, completeness, temporal resolution, accessibility, and relationship to a specific clinical decision. Moreover, information in the clinical record reflects not only patient states and clinical reasoning, but also documentation practices, institutional workflows, measurement processes, care fragmentation, and the circumstances under which data are collected. Embedding such knowledge in LLMs is therefore not simply a matter of supplying more text or improving answer accuracy. It requires methods for representing and linking heterogeneous evidence; retrieving, prioritizing, and updating relevant information; preserving provenance and temporal context; identifying uncertainty, missingness, and disagreement among sources; and evaluating whether model outputs, supporting evidence, and intermediate representations are clinically meaningful, well calibrated, and reliable across settings.

These requirements are central to developing medical LLMs that can support, rather than obscure, responsible clinical decision-making.

The PSB 2027 session, "NLP Methods for Embedding Real-World Clinical Knowledge in LLMs: Towards Responsible, Transformative Medical AI," was organized around these questions. Building on PSB's longstanding role in bringing together biomedical NLP, clinical informatics, knowledge representation, and translational biomedical research, the session sought work that treated medical knowledge as more than a static collection of facts. Our intent was to examine how clinical knowledge can be represented, grounded, updated, evaluated, and governed inside or alongside LLM-based systems. We were particularly interested in approaches that connected methodological innovation to real clinical data, patient-centered information, expert judgment, safety, fairness, and deployment constraints. The seven accepted papers collectively realize this agenda, although they approach it from different points in the knowledge pipeline.

2. Accepted papers

The accepted papers span clinical surveillance, patient-reported outcomes, longitudinal EHR reconstruction, synthetic clinical trajectories, medical search, evidence synthesis, and diagnostic reasoning. They illustrate two complementary directions: making clinical information more explicit and usable for downstream reasoning, and evaluating or training models to retrieve, justify, and align their answers with clinical evidence. The papers are summarized below in that order.

Weissenbacher and colleagues from Cedars-Sinai Medical Center, Cedars-Sinai Health System, and the University of Pennsylvania develop LLM-based systems for identifying in-hospital cardiac arrest events and their locations from routine EHR notes. The study compares guideline-enriched prompting, supervised fine-tuning, and self-correction against ICD-code-based extraction. Fine-tuning substantially improves performance on a curated corpus, while a two-pass correction workflow is more robust on a large, unselected validation cohort. The practical contribution is a scalable approach for prioritizing encounters for clinician adjudication and supporting cardiac-arrest surveillance and registry development¹.

Habibi and colleagues from Emory University School of Medicine introduce an observability-aware framework for evaluating language-based clinical phenotyping when the reference standard and model input come from different measurement channels. Using patient-reported depression outcomes, the study distinguishes model-recognition failure from limitations in what the analyzed clinical text can observe. The addition of patient portal messages improves sensitivity, while content-based abstention mainly enriches the positive population rather than reliably reducing selective risk. The work provides a way to interpret apparent model errors in light of missing or unrecorded patient evidence².

Kumar and colleagues from the National Institutes of Health and Carnegie Mellon University formulate clinical timeline reconstruction as retrieval-augmented temporal grounding. Narrative text supplies event semantics, while structured EHR rows provide partial temporal evidence from labs, medications, vitals, and procedures. Their scaffolded workflow extracts central events, builds an initial timeline, attaches additional events, and calibrates timestamps using structured data. Multimodal alignment generally improves temporal performance, although uncertainty remains

substantial and many text-derived events have no structured counterpart. The paper shows both the value and the limits of treating structured EHR data as temporal evidence³.

Zhou and colleagues from Stanford University, the University of Texas Dell Medical School, the University of Oklahoma Health Sciences Center, and LSU Health Shreveport introduce Monopath Directed Acyclic Graphs, which convert clinical case reports into ordered patient states linked by explicit clinical transitions. Using lung-cancer case reports, they evaluate graph fidelity, embedding structure, and graph-conditioned synthetic-case generation. Physicians rate graph-conditioned cases as more temporally valid, actionable, and safe for decision-making than text-only generations, although clinical validity itself does not significantly differ. The contribution is a structured intermediate representation for preserving decision-relevant clinical trajectories, especially in rare or atypical cases⁴.

Zhan and colleagues from the University of Minnesota introduce MedSR, a framework that trains compact language models to generate auditable reason-search-answer trajectories. The method combines an external search tool, LoRA adaptation, group relative policy optimization, and rule-based rewards without a separately trained reward model. Across medical question-answering benchmarks and retrieval corpora, learned search behavior can outperform untrained retrieval prompting, but the benefit depends on the model, corpus, retrieval depth, reward design, and adapter capacity. The paper treats search as a learned component of medical reasoning rather than a fixed preprocessing step⁵.

Liu and colleagues from the National Institutes of Health and the University of Illinois Urbana-Champaign evaluate whether contemporary chatbots retrieve the same primary clinical studies selected by Cochrane reviewers. Across models, user roles, and repeated queries, retrieval recall varies substantially, with researcher-oriented prompts performing better than clinician- or patient-oriented prompts. The analysis also identifies a bias toward larger clinical trials, even after accounting for publication year, citations, and open-access status. The study contributes an evidence-retrieval benchmark that exposes selection and coverage bias, extending evaluation beyond citation fabrication⁶.

Wu and colleagues from Stanford University, the University of Southern California, and the University of California, San Francisco introduce MedCaseReasoning, an open-access dataset of nearly 14,000 diagnostic question-answer cases derived from clinical case reports and paired with explicit reasoning statements. The authors show that current reasoning models have important deficits in both diagnostic accuracy and recall of documented reasoning, and that fine-tuning on the reasoning traces improves both measures. The contribution is an evaluation and training resource that treats faithful diagnostic reasoning, rather than only the final diagnosis, as a core dimension of medical LLM performance⁷.

3. Discussion

Several points of convergence emerge across the papers. First, the accepted work moves toward explicit and auditable knowledge representations. Timelines, state-transition graphs, reasoning traces, search trajectories, and observability annotations expose information that would otherwise remain implicit in narrative text or latent in model parameters. These representations do not

eliminate ambiguity, but they make it possible to ask more precise questions about provenance, temporal order, missing evidence, and the relationship between a model's intermediate behavior and its final output.

Second, the papers treat clinical knowledge as heterogeneous and complementary. Narrative notes can describe phenotype and clinical meaning; structured records can anchor time; patient-reported information can reveal symptoms that were not documented by clinicians; case reports can preserve rare trajectories; and the biomedical literature can supply evidence beyond the local record. The resulting systems are strongest when they combine these sources without assuming that any one source is complete. This also means that retrieval and representation choices are inseparable from questions of access, selection, missingness, and bias.

Third, the papers collectively argue for evaluation beyond end-task accuracy. Relevant questions include whether a patient state was temporally reconstructed, whether evidence was observable in the input, whether retrieved studies align with expert selection, whether a reasoning trace reflects clinically meaningful considerations, and whether generated content supports a safe next action. Human review remains important across these settings, but the papers also show the need for structured and reproducible evaluation protocols. Performance on curated datasets can differ from performance in broad, unselected clinical populations, as illustrated by the cardiac-arrest validation cohort, the observability analysis of undocumented symptoms, and the retrieval study's dependence on user role and study characteristics.

The collection also identifies important gaps. Most studies remain single-task or single-domain evaluations, leaving open questions about transportability across institutions, populations, specialties, and documentation practices. The field lacks shared standards for representing provenance, uncertainty, temporal validity, and conflicts among sources. Patient-generated knowledge is represented in one important study, yet systematic methods for incorporating patient priorities, community narratives, and lived experience into model behavior remain underdeveloped. The session proposal also emphasized continual updating, fairness, governance, and deployment monitoring, while the accepted papers provide only an initial view of how these requirements can be operationalized. Prospective evaluation of human-AI teams, workflow effects, escalation behavior, and longitudinal model drift is a clear next step.

These convergences suggest several opportunities. Shared benchmarks could combine evidence retrieval, longitudinal reconstruction, diagnostic reasoning, observability, and safety-oriented judgment rather than treating them as isolated tasks. Such benchmarks should include multiple information channels and explicitly record what evidence was available, retrieved, omitted, or contradicted. Cross-institutional studies could test whether structured clinical knowledge representations generalize across documentation cultures and patient populations. Evaluation suites should pair automated measures with expert assessment of actionability, calibration, fairness, and safe failure. Finally, knowledge-aware systems need governance mechanisms that make updates, source provenance, model limitations, and escalation paths visible to the people responsible for deploying and using them.

4. Conclusion

Taken together, the seven papers mark a transition from medical LLMs that merely produce answers toward systems that can expose, retrieve, structure, justify, and be evaluated against the clinical knowledge behind those answers. The session's central insight is that responsible medical LLM development requires treating knowledge representation, evidence access, reasoning behavior, and evaluation as a connected design problem. The next phase of work will require connecting these components into robust, continuously monitored, and clinically accountable systems. By bringing together methods for knowledge representation, retrieval, reasoning, evaluation, and real-world implementation, this session aims to help define what responsible and transformative medical LLMs should know, how they should demonstrate that knowledge, and how the field should govern their use in practice.

References

1. Weissenbacher D, Vo J, Uy-Evanado A, Reinier K, Chugh H, Vishnu K, Venkatesan M, O'Connor K, Chugh SS, Gonzalez-Hernandez G. Beyond ICD Codes: Fine-Tuning LLMs for In-Hospital Cardiac Arrest Identification from EHR Notes. In: Pacific Symposium on Biocomputing 2027.
2. Habibi S, Kavalieratos D, Parikh RB, Bozkurt S. Learning from Imperfect Patient-Reported Outcomes: An Observability-Aware Framework for Language-Based Clinical Phenotyping. In: Pacific Symposium on Biocomputing 2027.
3. Kumar S, Noroozizadeh S, Kim J, Weiss JC. Text Knows What, Tables Know When: Clinical Timeline Reconstruction via Retrieval-Augmented Multimodal Alignment. In: Pacific Symposium on Biocomputing 2027.
4. Zhou A, Fanous A, Bikia V, Deng Y, Natarajan S, Xu S, Agarwal AA, Garg A, Fanous N, Huang L, Tolbert P, Luu J, Alsentzer E, Daneshjou R. Monopath DAGs: Structured Patient Trajectories from Clinical Case Reports. In: Pacific Symposium on Biocomputing 2027.
5. Zhan Z, Wang J, Song Y, Zeng M, Yu K, Cui M, Lin M, Xie F, Zhang R. MedSR: Aligning Language Models to Reason and Search for Medical Question Answering with Group Relative Policy Optimization. In: Pacific Symposium on Biocomputing 2027.
6. Liu Q, Jin Q, Menke JD, Kahnt T, Lu Z. Do AI Chatbots Find What Experts Would? Effects of Model, User Role, and Sample Size on Study Retrieval for Medical Questions. In: Pacific Symposium on Biocomputing 2027.
7. Wu K, Wu E, Thapa R, Wei K, Zhang A, Suresh A, Tao JJ, Sun MW, Lozano A, Zou J. MedCaseReasoning: Evaluating and Learning Diagnostic Reasoning from Clinical Case Reports. In: Pacific Symposium on Biocomputing 2027.

Learning from Imperfect Patient-Reported Outcomes: An Observability-Aware Framework for Language-Based Clinical Phenotyping

Sheida Habibi¹, Dio Kavalieratos², Ravi B. Parikh², and Selen Bozkurt¹

¹*Biomedical Informatics Department, Emory University, Atlanta, GA 30322, USA*

²*Emory University School of Medicine, Atlanta, GA 30322, USA*

E-mail: sheida.habibi@emory.edu

Language models applied to clinical text are often evaluated against diagnosis codes or chart review, which reflect what clinicians recognized and documented rather than the full patient experience. Independently collected patient-reported outcomes reduce this dependence but create a challenge: when the reference standard and model input arise from different measurement channels, apparent errors may reflect either failure to recognize available evidence or absence of relevant evidence from the analyzed text. We introduce an observability-aware framework for evaluating clinical language models under this partial information. We demonstrate the framework in 1,725 palliative care patients without a prior depression diagnosis or antidepressant treatment, using Patient Health Questionnaire-8 (PHQ-8) scores ≥ 5 as the reference standard. Llama 3.3 70B Instruct classified clinical notes and portal messages, with predictions combined by logical OR. Clinical notes achieved sensitivity 0.67, specificity 0.58, and positive-class F1 0.64. Adding portal messages increased sensitivity to 0.71 and F1 to 0.65, recovering 34 additional positive patients while reducing specificity by 0.03. We defined lexicon-relative observability as the presence of at least one depressive-symptom domain in clinical notes. Approximately 80% of PHQ-8-positive patients were observable, with sensitivity 0.76 in this subgroup. Blinded review identified clinically meaningful evidence in 72.9% of lexicon-unobservable true positives versus 12.5% of lexicon-unobservable false negatives, with strong agreement ($\kappa = 0.85$). Content-conditioned abstention increased sensitivity and F1 but did not consistently reduce selective risk, indicating case enrichment rather than improved reliability. Observability-aware evaluation can help determine whether improvements should focus on stronger models, broader text sources, or direct outcome collection.

Keywords: clinical natural language processing; large language models; patient-reported outcomes; depression; palliative care.

1. Introduction

Depression is a leading cause of disability and is associated with reduced quality of life, increased healthcare utilization, suicide risk, and premature mortality.^{1,2} Despite effective treatments, depressive symptoms remain frequently underrecognized because of stigma, limited access to mental health services, brief encounters, and reluctance to disclose emotional concerns.³⁻⁵ This challenge is consequential in palliative care, where mood symptoms may be overlooked amid serious illness or difficult to distinguish from fatigue, sleep disturbance, appetite changes, and functional decline resulting from the underlying illness.^{21,22}

Electronic health records (EHRs) offer an opportunity to identify depressive symptoms during routine care. EHR phenotyping based on diagnosis codes, medications, and billing data is useful for established disease but less effective before symptoms are formally recognized.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

The authors have no conflicts of interest relevant to this article to disclose. This work was supported in part by a philanthropic gift from the Fraser Parker Foundation to Emory University's Division of Palliative Medicine. Dr. Bozkurt's effort was supported in part by the National Institute of Mental Health of the National Institutes of Health under award number R21MH139049.

Unstructured clinical text, including clinical notes, patients' portal messages and ambient transcripts of visits, can capture mood changes, emotional distress, psychosocial stressors, sleep problems, and functional impairment that precede diagnosis.^{6–9} Advances in natural language processing (NLP), including transformer-based models and large language models (LLMs), have therefore enabled sophisticated depression phenotyping from clinical documentation.^{10–15} These studies, however, have mostly relied on diagnosis codes, chart review, or expert-assigned labels as the reference standard. Such labels introduce a dependency that is analytically consequential: the reference standard and the model input derive from the same clinical encounter and the same clinician's judgment. Apparent model performance may therefore reflect the detection of documented clinical recognition rather than the patient symptoms, and the two are not interchangeable. Recognition is largely a precondition for coding, so patients whose symptoms are already recognized are, by construction, the patients for whom automated identification offers the least incremental value. Conversely, patients with unrecognized symptoms constitute the population of principal clinical interest, and this population cannot be operationally defined using a label derived from clinician recognition.

Independently collected patient-reported outcomes break this dependence. The Patient Health Questionnaire-8 (PHQ-8) measures symptom frequency over a defined recall period, whereas clinical text contains only information expressed, elicited, and documented within the available window. A patient may therefore meet the outcome threshold without corresponding depressive-symptom evidence in the model input. Conventional sensitivity treats such cases identically to those in which relevant evidence is present but missed, conflating limited textual observability with model-recognition failure. Thus, the patient-reported reference is not necessarily invalid but may be partially observable from the language modality being evaluated.

The text source further shapes what becomes observable. Clinical notes reflect information elicited during encounters and clinical interpretation but may omit concerns that patients do not disclose or clinicians do not document. Patient portal messages provide a complementary patient-generated channel for describing symptoms, emotional experiences, treatment concerns, and requests for support between visits. Prior studies have demonstrated their utility for identifying medication-related issues, adverse events, and mental health concerns, including depression-related content.^{16–20} However, their incremental contribution beyond clinical notes remains insufficiently characterized.

Existing language-based phenotyping studies generally emphasize full-cohort discrimination without assessing if phenotype-relevant evidence is present in the analyzed text.^{10–15} Few approaches quantify text-relative observability or determine whether restricting prediction to content-rich records reduces error or merely enriches the retained population for positive cases.

In this study we develop and apply an observability-aware evaluation of language-based depressive symptom detection against an independently collected patient-reported outcome. The analysis proceeds in three parts. First, we benchmark an instruction-tuned LLM applied to clinical notes and patient-generated portal messages and interpret its performance in relation to trivial, lexical, and supervised baselines commonly used in clinical phenotyping. Second, we partition outcome-positive patients by model prediction and by lexicon-observability, defined as the presence of explicit depressive symptom language in the analyzed text, which

separates cases missed despite available evidence from cases in which no such evidence was documented. We assessed the validity of this grouping through a blinded review of lexicon-unobservable cases. Third, we quantified the additional contribution of portal messages and, in a separate clinical-note-only analysis, evaluated whether the number of lexicon-detected symptom domains could serve as an abstention criterion across different coverage levels. We release the prompts, lexicon, and analysis code to support replication.

2. Methodology

Figure 1 summarizes the study workflow.

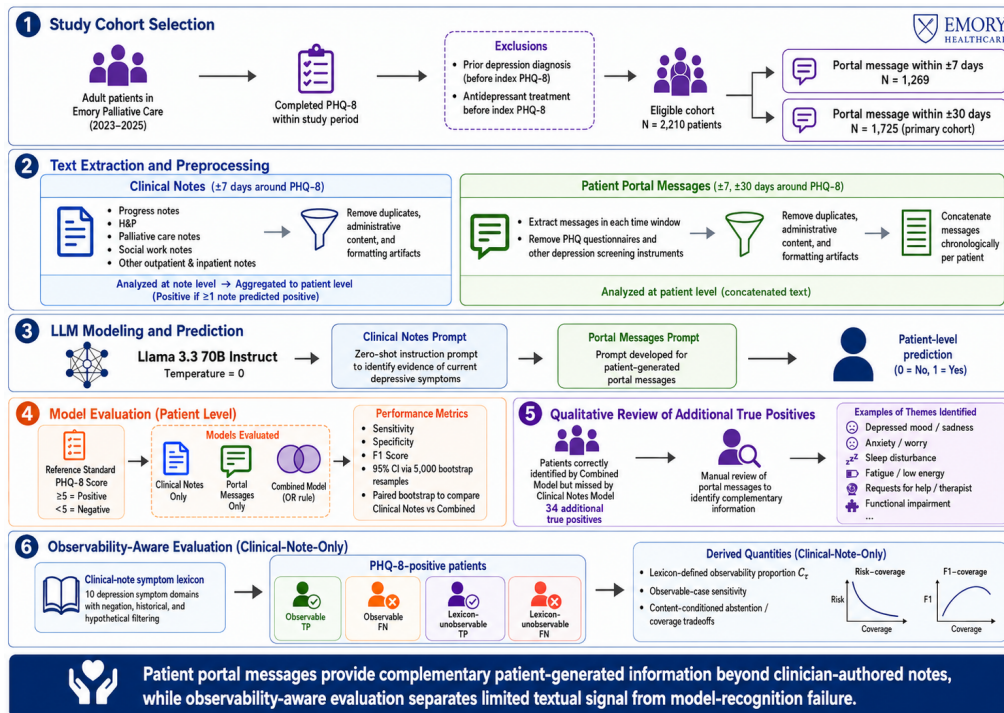


Fig. 1. Study workflow, including cohort selection, text preprocessing, LLM-based prediction, patient-level evaluation, and observability-aware error decomposition.

2.1. Study design and cohort

We included adult patients who received palliative care within Emory Healthcare and completed at least one PHQ-8 assessment between January 1, 2023, and December 31, 2025. The study was approved by the Emory University Institutional Review Board. Eligible patients had at least one PHQ-8 and at least one clinical note within ± 7 days. For patients with multiple eligible assessments, the first was the index assessment.

To focus on unrecognized symptoms and to prevent the model from reading a diagnosis off the chart, we excluded 976 patients with a documented depression diagnosis (ICD-10-CM) or antidepressant exposure before the index PHQ-8, leaving 2,210 eligible patients. Among these, 1,269 had at least one portal message within ± 7 days of the index assessment and 1,725 within ± 30 days. The ± 30 -day portal-message cohort was designated as the primary analytic cohort

because it retained more eligible patients and therefore provided greater statistical precision, whereas the ± 7 -day portal-message cohort was used as a temporal sensitivity analysis.

2.2. *Reference Standard*

The independently collected PHQ-8 assesses eight depressive symptoms over the preceding two weeks, with items scored from 0 to 3 and total scores ranging from 0 to 24. Scores of 0–4 were classified as minimal or no symptoms and scores ≥ 5 as at least mild symptom burden, a threshold selected to prioritize sensitivity for early identification rather than clinical diagnosis.^{23,24} Because the PHQ-8 and analyzed text represented distinct measurement channels, patients could meet the threshold without corresponding textual evidence or describe distress despite scoring below it. Performance therefore depended on both textual evidence availability and model recognition.

2.3. *Text Processing and Model Configuration*

We extracted clinical notes and patient-generated MyChart portal messages within the pre-specified temporal windows. The symmetric ± 7 -day clinical-note window accommodated documentation that may be entered or finalized after the corresponding encounter while maintaining temporal proximity to the index assessment. Clinical notes included progress notes, palliative care notes, social work notes, and other documentation. Portal messages were ordered chronologically and concatenated into a patient-level document for each temporal window.

To minimize information leakage, embedded PHQ questionnaires and other structured depression-screening instruments were removed before inference, along with duplicate records, administrative or automated content, and formatting artifacts. Depressive symptoms were then classified using Llama 3.3 70B Instruct without task-specific fine-tuning,²⁵ with temperature set to 0. Separate zero-shot prompts were developed for clinical notes and patient-generated portal messages. Text was classified as positive only when it contained evidence of current depressive symptoms; historical, hypothetical, or negated symptoms and antidepressant use without accompanying symptom evidence were insufficient for a positive classification.

We evaluated three patient-level prediction strategies: (1) clinical notes only, in which a patient was classified as positive if at least one note received a positive prediction; (2) portal messages only, in which a prediction was generated from the concatenated patient-level portal-message document; and (3) combined text sources, in which clinical note and portal message predictions were integrated using a logical OR rule. The combined strategy assessed whether patient-generated communication contributed information beyond clinical documentation.

2.4. *Manual Review of Complementary Portal-Message Cases*

We manually reviewed all additional PHQ-8-positive patients identified by the combined OR strategy but missed by the clinical-note-only model. Portal messages were compared with corresponding clinical notes to identify complementary information. Non-mutually exclusive themes were summarized as counts, and one de-identified representative portal-message excerpt was reported for each theme.

2.5. *Lexicon-Based Measurement of Textual Observability*

To distinguish absent detected evidence from recognition failure, we defined a text-only, label-independent proxy for whether each patient's notes contained explicit depressive-symptom content. The lexicon was built from established symptom constructs and documentation terminology, covering 10 domains: depressed mood, anhedonia, sleep disturbance, fatigue or low energy, appetite or weight change, worthlessness or guilt, impaired concentration, psychomotor change, mental-health help-seeking, and suicidality or self-harm. Terms and regular expressions were developed without access to PHQ-8 labels or model predictions. Mentions were excluded when preceded within a fixed left-context window by negation cues such as "denies," historical cues such as "history of," or hypothetical cues such as "monitor for".^{26,27}

For patient i , c_i denoted the number of distinct symptom domains remaining after contextual filtering. Observability at threshold τ was defined as

$$O_i(\tau) = \begin{cases} 1, & \text{if } c_i \geq \tau, \\ 0, & \text{if } c_i < \tau. \end{cases} \quad (1)$$

Patients were lexicon-observable when at least τ domains were detected. The indicator was calculated from text alone and did not use the PHQ-8 label or model prediction.

Observability was lexicon-relative meaning notes without detected symptom terms could still contain clinically meaningful evidence of depression. Observability was assessed only in clinical notes from the 1,725-patient primary cohort.

2.6. *Decomposition of Outcome-Positive Cases*

Observability was used as an explanatory stratification and did not affect cohort inclusion. All 1,725 patients remained in full cohort performance estimates and selective prediction analyses.

Of the P PHQ-8-positive patients, each was assigned to one of four mutually exclusive groups defined jointly by the clinical-note prediction $\hat{y}_i$ and the observability indicator $O_i(\tau)$:

$$\begin{aligned} TP_{\text{obs}} &= \sum_{i=1}^N \mathbb{1}(y_i = 1, \hat{y}_i = 1, O_i(\tau) = 1) & FN_{\text{obs}} &= \sum_{i=1}^N \mathbb{1}(y_i = 1, \hat{y}_i = 0, O_i(\tau) = 1) \\ TP_{\text{unobs}} &= \sum_{i=1}^N \mathbb{1}(y_i = 1, \hat{y}_i = 1, O_i(\tau) = 0) & FN_{\text{unobs}} &= \sum_{i=1}^N \mathbb{1}(y_i = 1, \hat{y}_i = 0, O_i(\tau) = 0). \end{aligned} \quad (2)$$

These groups satisfied:

$$P = TP_{\text{obs}} + FN_{\text{obs}} + TP_{\text{unobs}} + FN_{\text{unobs}}. \quad (3)$$

Here, FN_{obs} represented missed cases despite detected depressive content, whereas FN_{unobs} represented missed cases without detected explicit content. The number of lexicon-unobservable positive cases was

$$U_{\tau} = TP_{\text{unobs}} + FN_{\text{unobs}} = \sum_{i=1}^N \mathbb{1}(y_i = 1, O_i(\tau) = 0). \quad (4)$$

Accordingly, the observability proportion can also be defined as

$$C_{\tau} = 1 - \frac{U_{\tau}}{P}. \quad (5)$$

Thus, C_τ represented the proportion of PHQ-8-positive patients whose notes contained at least τ lexicon-detected symptom domains. For methods relying only on the explicit evidence captured by the lexicon, however, it could be interpreted as a lexicon-relative upper bound.

Sensitivity among lexicon-observable positive cases was compared descriptively with full-cohort sensitivity, computed as

$$\text{Sensitivity}_{\text{obs}} = \frac{TP_{\text{obs}}}{TP_{\text{obs}} + FN_{\text{obs}}}, \quad \text{Sensitivity} = \frac{TP}{TP + FN}. \quad (6)$$

This comparison characterizes performance within the lexicon-defined observable subset but does not estimate sensitivity loss caused by absent textual evidence. We reported TP_{unobs} separately because positive predictions in lexicon-unobservable cases may reflect implicit evidence missed by the lexicon, nonspecific correlates, or unsupported predictions.

2.7. Sensitivity Analysis of the Observability Definition

We repeated the positive-case decomposition using observability thresholds of $\tau = 1, 2, 3$ and 4, corresponding to detection of at least one, two, three or four distinct symptom domains. At each threshold, we recalculated the four observability subgroups, U_τ , C_τ , and $\text{Sensitivity}_{\text{obs}}$.

The primary definition used $\tau = 1$ because it was the most inclusive. Stricter thresholds were evaluated only to assess sensitivity to the operational definition and were not interpreted as more accurate estimates of semantic observability, because increasing τ deliberately reclassified some patients with detected content as lexicon-unobservable.

2.8. Content-Conditioned Abstention and Selective Prediction

We evaluated the content count c_i as an input-side abstention criterion. At threshold t , patients with $c_i \geq t$ were retained, regardless of PHQ-8 label, and patients with $c_i < t$ were excluded only from performance calculations. The abstention rule used neither the PHQ-8 label, the model prediction, nor model correctness. Unlike conventional selective-classification approaches, in which rejection is typically based on model confidence or uncertainty,^{28,29} our retention rule was based solely on the amount of phenotype-relevant content available in the input. We nevertheless adopted the standard risk-coverage formalism from selective prediction.^{28,29}

The retained set and coverage were defined as

$$R(t) = \{i : c_i \geq t\}, \quad \text{Coverage}(t) = \frac{|R(t)|}{N}, \quad (7)$$

where N was the size of the primary analytic cohort.

Selective risk, defined as the proportion of incorrect predictions among patients retained at threshold t , was calculated under 0–1 loss as

$$\text{Risk}(t) = 1 - \frac{TP(t) + TN(t)}{|R(t)|}. \quad (8)$$

Within each retained set, we calculated sensitivity, specificity, precision, positive-class F_1 score, and PHQ-8-positive prevalence, p , which represented the proportion of retained patients with $\text{PHQ-8} \geq 5$ and defined as

$$p(t) = \frac{\sum_{i \in R(t)} \mathbb{1}(y_i = 1)}{|R(t)|}. \quad (9)$$

Because retained-set prevalence could vary across thresholds, selective risks were not interpreted in isolation. At each threshold, we also calculated the risk of the optimal trivial constant classifier, that predicted the majority class for every retained patient, as

$$\text{Risk}_{\text{trivial}}(t) = \min \{p(t), 1 - p(t)\}. \quad (10)$$

The model’s prevalence-adjusted advantage over trivial classification was then defined as

$$\Delta_{\text{risk}}(t) = \text{Risk}_{\text{trivial}}(t) - \text{Risk}(t). \quad (11)$$

Positive values indicated lower error than majority-class prediction in the retained set.

Thresholds were evaluated across c_i values, excluding those retaining fewer than 20 patients. Risk–coverage, F_1 –coverage, and risk-advantage–coverage curves assessed performance across thresholds. Abstention was considered reliability-enhancing only if risk decreased while advantage over the threshold-specific trivial classifier was maintained or increased; otherwise, gains in sensitivity or F_1 were interpreted as case enrichment or screening prioritization.

2.9. Performance Evaluation and Statistical Analysis

All analyses were conducted at the patient level. Notes-only, portal-only, and combined strategies were evaluated using accuracy, sensitivity, specificity, precision, positive-class F_1 , and macro F_1 ; abstention analyses additionally reported coverage, selective risk, and retained-set prevalence. Ninety-five percent confidence intervals were estimated using 5,000 patient-level bootstrap resamples. Differences between strategies were assessed using 5,000 paired bootstrap resamples that preserved both predictions within each resample, with two-sided bootstrap p -values calculated for differences in sensitivity, specificity, and positive-class F_1 .³⁰

2.10. Blinded Semantic Audit of Lexicon-Unobservable Cases

To assess the semantic validity of lexicon-unobservable cases, two reviewers independently evaluated all 48 true positives and a random sample of 48 lexicon-unobservable false negatives, blinded to PHQ-8 scores, LLM predictions, lexicon output, and subgroup membership. Cases were categorized as containing clinically meaningful, ambiguous or indirect, or no evidence of current depressive symptoms. Disagreements were resolved by consensus, and inter-rater agreement was assessed using Cohen’s κ . The audit was designed to assess the unobservable designation among PHQ-8-positive cases; it does not establish the validity of the full four-cell decomposition.

3. Results

Table 1 reports cohort characteristics overall and by PHQ-8 status.

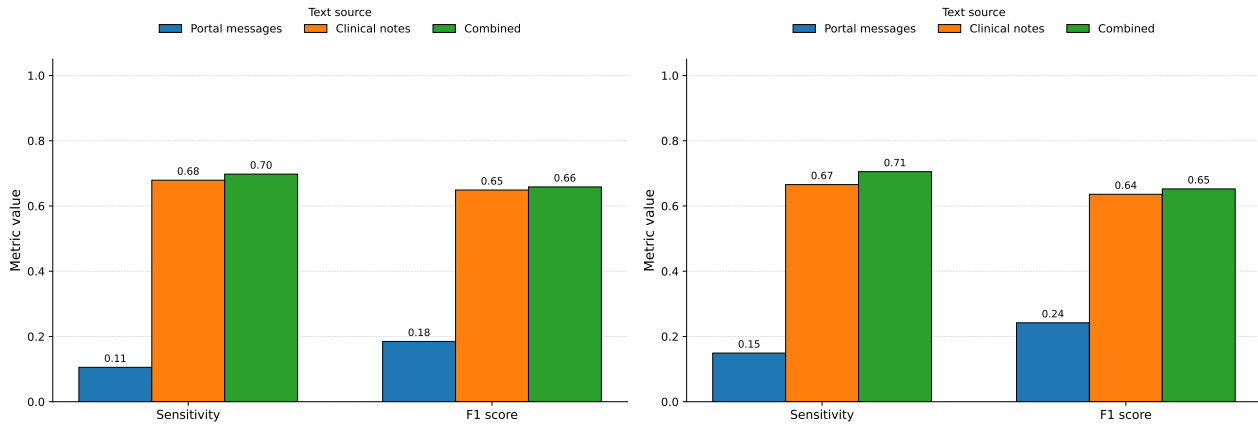
3.1. Model Performance

Figure 2 summarizes performance across both portal-message windows. Clinical notes outperformed portal messages alone, while combining sources increased sensitivity and positive-class F_1 but reduced specificity (Table 2). In the primary ± 30 -day cohort, sensitivity rose from 0.67 to 0.71 and F_1 from 0.64 to 0.65, while specificity fell from 0.58 to 0.55.

Table 1. Characteristics of the primary analytic cohort overall and stratified by PHQ-8 status.

Characteristic	Overall (<i>N</i> = 1,725)	PHQ-8 < 5 (<i>N</i> = 867)	PHQ-8 ≥ 5 (<i>N</i> = 858)
Age, years	60.8 (14.7)	61.0 (14.5)	60.7 (15.0)
Clinical-note count per patient	13.1 (15.2)	12.6 (16.6)	13.6 (13.6)
Portal-message count per patient	26.9 (26.5)	25.2 (25.1)	28.6 (27.8)
Gender			
Female	1,028 (59.6%)	505 (58.2%)	523 (61.0%)
Male	697 (40.4%)	362 (41.8%)	335 (39.0%)
Race			
American Indian or Alaska Native	5 (0.3%)	2 (0.2%)	3 (0.3%)
Asian	73 (4.2%)	28 (3.2%)	45 (5.2%)
Black or African American	609 (35.3%)	326 (37.6%)	283 (33.0%)
Native Hawaiian/Pacific Islander	2 (0.1%)	2 (0.2%)	0 (0.0%)
White	957 (55.5%)	479 (55.2%)	478 (55.7%)
More than one race	9 (0.5%)	6 (0.7%)	3 (0.3%)
Other/unknown	70 (4.1%)	24 (2.8%)	46 (5.4%)
Ethnicity			
Hispanic or Latino	83 (4.8%)	37 (4.3%)	46 (5.4%)
Not Hispanic or Latino	1,585 (91.9%)	805 (92.8%)	780 (90.9%)
Unknown or declined	57 (3.3%)	25 (2.9%)	32 (3.7%)

Values are mean (SD) for continuous variables and *n* (%) for categorical variables.



(a) Using ±7-Day Portal-Message Window (b) Using ±30-Day Portal-Message Window

Fig. 2. Model Performance Comparison for Clinical Notes, Portal Messages, and Combined Sources

3.2. Portal Messages Contribution to Additional True-Positive Detection

Among the 34 additional PHQ-8-positive cases, portal messages most commonly described depressed mood (14 cases), anxiety or emotional distress (13), and fatigue or low energy (10) (Table 3). Compared with notes, which often emphasized physical symptoms, treatment, or disease management, portal messages contained more explicit psychosocial information.

3.3. Lexicon-Defined Observability and Positive-Case Decomposition

At the primary threshold of $\tau = 1$, 685 of the 858 PHQ-8-positive patients met the lexicon-defined observability criterion, yielding an observability proportion of $C_{\tau=1} = 0.80$. The remain-

Table 2. Patient-level performance of the clinical-note-only and combined OR strategies across portal-message windows.

±7-Day Portal-Message Window				
Metric	Clinical notes	Combined	Difference with 95% CI	Bootstrap p-value
Accuracy	0.63	0.63	0.006 (-0.001 to 0.012)	0.114
Precision	0.62	0.62	0.002 (-0.004 to 0.007)	0.456
Sensitivity	0.68	0.70	0.019 (0.009 to 0.030)	< 0.001
Specificity	0.57	0.56	-0.008 (-0.016 to -0.002)	0.010
F_1 score	0.65	0.66	0.009 (0.003 to 0.017)	0.004
Macro F_1	0.62	0.63	0.005 (-0.001 to 0.011)	0.129
±30-Day Portal-Message Window				
Metric	Clinical notes	Combined	Difference with 95% CI	Bootstrap p-value
Accuracy	0.62	0.63	0.005 (-0.003 to 0.014)	0.288
Precision	0.61	0.61	-0.002 (-0.010 to 0.006)	0.599
Sensitivity	0.67	0.71	0.040 (0.027 to 0.053)	< 0.001
Specificity	0.58	0.55	-0.03 (-0.040 to -0.019)	< 0.001
F_1 score	0.64	0.65	0.016 (0.007 to 0.026)	< 0.001
Macro F_1	0.62	0.62	0.004 (-0.005 to 0.012)	0.415

Table 3. Themes identified during review of portal messages from the 34 additional PHQ-8-positive patients.

Type of Information	Count	Short Portal Example (De-identified)
Depressed mood	14	“I’m getting extremely depressed and looking for solutions”
Emotional distress	13	“Unfortunately I’ve been having panic attacks and heightened anxiety lately... with nothing to combat it with.”
Fatigue or low energy	10	“I feel fatigued. I slept all weekend and that is not like myself.”
Explicit request for mental health help	9	“I need a therapist to talk to about my cancer and being terminally ill. Can you please put me in touch with someone?”
Sleep disturbance	8	“I haven’t slept in 4 nights. This has never happened before.”
Appetite / weight changes	7	“I have a loss of appetite and barely eat.”
Feelings of hopelessness or worthlessness	3	“I have no plan from you... I passed around like a rag doll.”
Loss of interest or pleasure (anhedonia)	1	“He has been anhedonic and mostly sitting around doing nothing... doesn’t seem to have the motivation or interest in doing anything.”
Suicidal thoughts or self-harm concerns	1	“It’s like I’m having so much anxiety that it’s crippling me and I just want to kill myself!”
Financial or occupational stress	1	“The bill is unexpected and I can’t afford to pay. I have already had choose between eating and paying bills.”

ing 173 (20.2%) patients had no lexicon-detected depressive symptom domains. Among the 685 lexicon-observable patients, the LLM correctly identified 523, corresponding to an observable-case sensitivity of 0.76. Of the model’s 287 false negatives, 125 were lexicon-unobservable, whereas 162 contained detectable depressive content but were missed by the model. The LLM also produced positive predictions for 48 PHQ-8-positive patients whose notes contained no

lexicon-detected symptom domains, as presented in Table 4.

3.4. Sensitivity to the Operational Definition of Observability

In addition to the primary decomposition at $\tau = 1$, we repeated the analysis using thresholds of two, three, and four detected symptom domains (Table 4). As the threshold increased, lexicon-defined observability decreased, whereas sensitivity among lexicon-observable patients increased because the subgroup was restricted to patients with more explicit symptom content. Correctly predicted PHQ-8-positive patients classified as lexicon-unobservable increased from 48 at $\tau = 1$ to 437 at $\tau = 4$. Thus, lower C_τ values at stricter thresholds partly reflected reclassification of patients with fewer detected domains rather than the absence of lexicon-detected depressive content. We retained $\tau = 1$ as the primary threshold because it provided the most inclusive definition of documented symptom content.

Table 4. Decomposition of PHQ-8-positive patients across lexicon-defined observability thresholds.

Threshold	$TP_{\text{obs}}/FN_{\text{obs}}$	$TP_{\text{unobs}}/FN_{\text{unobs}}$	Unobservable positives	C_τ	Sensitivity _{obs}
≥ 1 domain	523 / 162	48 / 125	173	0.80	0.76
≥ 2 domains	415 / 84	156 / 203	359	0.58	0.83
≥ 3 domains	286 / 42	285 / 245	530	0.38	0.87
≥ 4 domains	134 / 12	437 / 275	712	0.17	0.92

3.5. Content-Conditioned Abstention and Coverage Tradeoffs

As shown in Figure 3, by increasing the content threshold, coverage decreased and the retained set became increasingly enriched for PHQ-8-positive patients. Sensitivity and positive-class F_1 improved, whereas specificity declined. However, selective risk did not consistently decrease, and the model’s advantage over the threshold-specific trivial classifier narrowed and became slightly negative at the lowest coverage. These findings indicate that content-conditioned abstention improved case detection primarily through class enrichment rather than through greater overall predictive reliability.

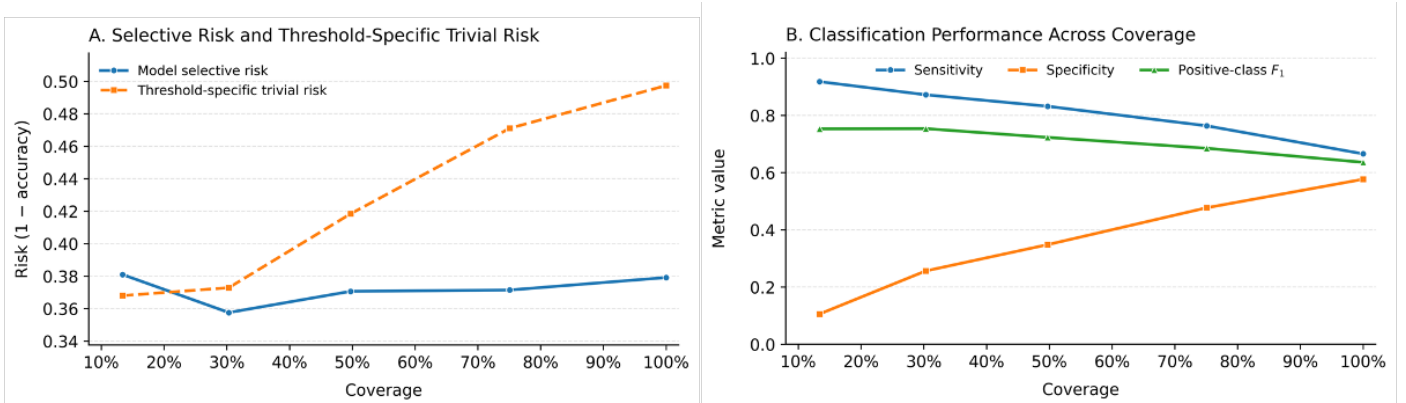


Fig. 3. Content-conditioned abstention across coverage levels. Panel A shows model and trivial-classifier risk; Panel B shows retained-set sensitivity, specificity, and positive-class F_1 .

3.6. Manual Validation of Lexicon-Unobservable Cases

In the blinded semantic audit, clinically meaningful depressive-symptom evidence was identified in 35 of 48 (72.9%) lexicon-unobservable true positives and 6 of 48 (12.5%) lexicon-unobservable false negatives. In true-positive cases, this evidence often consisted of indirect or contextual expressions not captured by the lexicon. Inter-rater agreement was strong ($\kappa = 0.85$). These findings indicate that lexicon-defined unobservability was instrument-relative: some notes contained meaningful symptom information beyond the lexicon, whereas cases negative by both the lexicon and model more often contained limited evidence.

4. Discussion

This study introduces a lexicon-based observability-aware framework that helps to distinguish failure to recognize available phenotype-relevant evidence from limited availability of that evidence in the analyzed text. We demonstrated the framework in palliative care, a setting in which depressive symptoms may overlap with serious illness and treatment effects, although the empirical estimates remain specific to this population and institution.

Conventional sensitivity treats all missed outcome-positive cases identically. Among the 287 PHQ-8-positive patients missed by the clinical-note model, 125 had no lexicon-detected depressive-symptom domains, whereas 162 contained lexicon-detected content but were still missed. Sensitivity was therefore higher among lexicon-observable positive cases than in the full cohort, 0.76 versus 0.67. This decomposition indicates whether improvement efforts should target model recognition or broader availability of phenotype-relevant evidence.

The blinded audit showed that the lexicon was informative but incomplete. Clinically meaningful depressive-symptom evidence was substantially more common among lexicon-unobservable true positives than false negatives, suggesting that the LLM recognized indirect or contextual evidence beyond explicit symptom terminology, whereas cases classified as negative by both the lexicon and model more often reflected limited documentation. Although the audit estimates remain imprecise, their confidence intervals did not overlap and inter-rater agreement was strong. Accordingly, $C_\tau = 0.80$ should be interpreted as a lexicon-defined observability proportion rather than a universal ceiling on text-based model performance.

Portal messages provided a complementary measurement channel. Adding them to clinical notes increased sensitivity from 0.67 to 0.71 and identified 34 additional PHQ-8-positive patients, often through psychosocial information that was absent from or insufficiently represented in the notes. However, specificity decreased from 0.58 to 0.55, suggesting that the logical OR strategy is better suited to screening and case prioritization than diagnostic classification.

Content-conditioned abstention did not consistently improve predictive reliability. Restricting evaluation to notes with more lexicon-detected symptom domains increased sensitivity and positive-class F1 but reduced specificity, enriched the retained population for PHQ-8-positive patients, and did not lower selective risk. These gains therefore reflected case enrichment rather than improved error control, making content-based restriction more useful for prioritization than reliability-enhancing rejection.

Observability-aware evaluation is applicable when reference standards and model inputs arise from different measurement channels, because conventional metrics alone cannot dis-

tinguish unavailable evidence from model-recognition failure. Studies using patient-reported outcomes, quality-of-life measures, or other independently collected labels should define and validate an input-specific observability instrument, jointly stratify outcome-positive cases by observability and model prediction, and report observable-subset performance alongside full-cohort performance. This decomposition can indicate whether improvements should focus on stronger models, broader or more appropriate text sources, or direct collection of the target outcome. Although the instrument must be adapted to each phenotype, setting, and text source, performance should be interpreted in relation to the information available to the model.

When structured or patient-reported outcomes are unavailable, agreement and disagreement between lexicon and model predictions may help prioritize patients for formal depressive-symptom screening or additional review. Concordant negative signals may indicate limited documented evidence but cannot exclude depressive symptoms. This text-only triage application was not evaluated and requires prospective validation.

Several limitations should be considered. Observability depended on the terms, symptom domains, and contextual filters included in the rule-based lexicon, and the semantic audit included only 48 lexicon-unobservable true positives and 48 lexicon-unobservable false negatives. Requiring portal-message availability may have introduced selection related to digital access, health literacy, healthcare engagement, and socioeconomic factors. Findings from this palliative-care population may not generalize to settings in which depressive symptoms are expressed or documented differently. Finally, this retrospective, single-system study evaluated one instruction-tuned LLM, prompting configuration, phenotype, and PHQ-8 threshold. Although the framework is independent of these choices, external and prospective studies are needed across institutions, models, text sources, populations, and phenotyping tasks.

5. Conclusion

Evaluation of language-based clinical phenotyping against independently collected patient-reported outcomes faces an interpretability problem: conventional sensitivity cannot distinguish cases missed despite relevant textual evidence from cases with little or no detectable evidence. Our observability-aware framework addresses this problem by stratifying outcome-positive patients by model prediction and lexicon-defined evidence availability. Portal messages broadened case detection by contributing patient-generated symptom information insufficiently represented in clinical notes. Content-conditioned abstention enriched the retained population for positive cases but did not reduce overall prediction error, supporting screening prioritization rather than reliability-enhancing rejection. Observability-aware analysis therefore provides a practical framework for interpreting language-based phenotyping when reference standards and model inputs arise from different measurement channels.

References

1. World Health Organization. Depression. Geneva, Switzerland: World Health Organization; 2023. Accessed July 22, 2026. <https://www.who.int/news-room/fact-sheets/detail/depression>.
2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Psychiatry*. 2022;9(2):137-150. doi:10.1016/S2215-0366(21)00395-3.
3. Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. *Lancet*. 2009;374(9690):609-619. doi:10.1016/S0140-6736(09)60879-5.
4. Clement S, Schauman O, Graham T, et al. What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychol Med*. 2015;45(1):11-27. doi:10.1017/S0033291714000129.
5. US Preventive Services Task Force. Screening for depression and suicide risk in adults: US Preventive Services Task Force recommendation statement. *JAMA*. 2023;329(23):2057-2067. doi:10.1001/jama.2023.9297.
6. Bozkurt S, Habibi S, Oltmann C, Pathak R, Kavalieratos D. Identifying depression in palliative care patients' messages using large language models. *J Pain Symptom Manage*. 2026;71(6):e902-e903.
7. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008;128-144. doi:10.1055/s-0038-1638592.
8. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*. 2012;13(6):395-405. doi:10.1038/nrg3208.
9. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural language processing for EHR-based computational phenotyping. *J Biomed Inform*. 2019;97:103272. doi:10.1016/j.jbi.2019.103272.
10. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Identifying symptom information in clinical notes using natural language processing. *Nurs Res*. 2021;70(3):173-183. doi:10.1097/NNR.0000000000000488.
11. Panaite V, Devendorf AR, Finch D, Bouayad L, Luther SL, Schultz SK. The value of extracting clinician-recorded affect for advancing clinical research on depression: proof-of-concept study applying natural language processing to electronic health records. *JMIR Form Res*. 2022;6(5):e34436. doi:10.2196/34436.
12. Leis A, Mayer MA, Gauna S, et al. Exploring the association of cancer and depression in electronic health records: combining encoded diagnosis and mining free-text clinical notes. *JMIR Cancer*. 2022;8(3):e39003. doi:10.2196/39003.
13. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2019:72-78. doi:10.18653/v1/W19-1909.
14. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. *arXiv*. Published online April 10, 2019. doi:10.48550/arXiv.1904.05342.
15. Grothman A, Ma WJ, Tickner KG, Martin EA, Southern DA, Quan H. Case identification of depression in inpatient electronic medical records: scoping review. *JMIR Med Inform*. 2024;12:e49781. doi:10.2196/49781.
16. Kruse CS, Argueta DA, Lopez L, Nair A. Patient and provider attitudes toward the use of patient portals for the management of chronic disease: a systematic review. *J Med Internet Res*. 2015;17(2):e40. doi:10.2196/jmir.3703.
17. van Buchem MM, de Hond AAH, Fanconi C, et al. Applying natural language processing to patient messages to identify depression concerns in cancer patients. *J Am Med Inform Assoc*.

- 2024;31(10):2255-2262. doi:10.1093/jamia/ocae188.
18. Duvall M, North F, Leasure W, Pecina J. Patient portal message characteristics and reported thoughts of self-harm and suicide: a retrospective cohort study. *J Telemed Telecare*. 2021;27(8):501-508. doi:10.1177/1357633X19887262.
 19. Huang M, Khurana A, Mastorakos G, et al. Patient portal messaging for asynchronous virtual care during the COVID-19 pandemic: retrospective analysis. *JMIR Hum Factors*. 2022;9(2):e35187. doi:10.2196/35187.
 20. Mermin-Bunnell K, Zhu Y, Hornback A, et al. Use of natural language processing of patient-initiated electronic health record messages to identify patients with COVID-19 infection. *JAMA Netw Open*. 2023;6(7):e2322299. doi:10.1001/jamanetworkopen.2023.22299.
 21. Lloyd-Williams M. Screening for depression in palliative care patients: a review. *Eur J Cancer Care (Engl)*. 2001;10(1):31-35.
 22. Noorani NH, Montagnini M. Recognizing depression in palliative care patients. *J Palliat Med*. 2007;10(2):458-464. doi:10.1089/jpm.2006.0099.
 23. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. *J Affect Disord*. 2009;114(1-3):163-173. doi:10.1016/j.jad.2008.06.026.
 24. Kroenke K, Spitzer RL, Williams JBW, Lowe B. The Patient Health Questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *Gen Hosp Psychiatry*. 2010;32(4):345-359. doi:10.1016/j.genhosppsych.2010.03.006.
 25. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models. *arXiv*. Published online July 31, 2024. doi:10.48550/arXiv.2407.21783.
 26. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 2001;34(5):301-310. doi:10.1006/jbin.2001.1029.
 27. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experienter, and temporal status from clinical reports. *J Biomed Inform*. 2009;42(5):839-851. doi:10.1016/j.jbi.2009.05.002.
 28. El-Yaniv R, Wiener Y. On the foundations of noise-free selective classification. *J Mach Learn Res*. 2010;11:1605-1641.
 29. Geifman Y, El-Yaniv R. Selective classification for deep neural networks. In: *Advances in Neural Information Processing Systems 30*. Curran Associates; 2017:4878-4887.
 30. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall/CRC; 1993.

Text Knows What, Tables Know When: Clinical Timeline Reconstruction via Retrieval-Augmented Multimodal Alignment

Sayantana Kumar^{1*}, Shahriar Noroozizadeh^{2,3*}, Juyong Kim^{2*}, Jeremy C. Weiss¹

¹*National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*

²*Machine Learning Department, Kenneth C. Griffin School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA*

³*Heinz College of Information Systems Management and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA*

Clinical language models increasingly operate over electronic health records (EHRs), yet patient records are not stored as temporally grounded trajectories. Clinical notes describe symptoms, assessments, and disease progression, but often compress or narratively reorder events. Structured EHR rows provide timestamps for labs, medications, vitals, and procedures, but capture only part of the clinical story. We formulate clinical timeline reconstruction as retrieval-augmented temporal grounding: constructing a patient trajectory by using narrative text for event semantics and structured rows as partial temporal evidence. We introduce a scaffolded workflow that extracts central narrative events, builds an initial temporal scaffold, attaches non-central events, and calibrates timestamps using retrieved structured EHR rows. We evaluate on 40 discharge summaries, including 15 i2b2-derived and 25 MIMIC-IV summaries, each with manual gold-standard timelines and aligned structured EHR data. Across models, multimodal calibration left event match rates largely unchanged and generally improved temporal performance: mean paired case-level multimodal–unimodal differences were positive in 7 of 12 model-metric comparisons across concordance and AULTC, with none negative. However, uncertainty was substantial given the 40-case sample; paired case-level bootstrap intervals excluded zero only for the DeepSeek V3.2 AULTC improvement. A gap analysis shows that 35.1% of text-derived events have no structured counterpart. These findings support treating structured EHR data as partial temporal evidence for narrative-derived patient trajectories.

Keywords: Clinical timeline reconstruction; electronic health records; clinical narratives; retrieval-augmented generation; multimodal alignment; large language models

1. Introduction

Time is central to clinical reasoning in critical care. Patients often follow rapidly evolving trajectories, requiring symptoms, physiologic measurements, treatments, and outcomes to be interpreted relative to admission, deterioration, intervention, and recovery. Sepsis trajectories exemplify this challenge. Although Sepsis-3 remains important for trial eligibility and computational phenotyping,^{1,2} sepsis is better understood as a set of heterogeneous clinical trajectories with a shared dysregulatory response rather than as a single, uniform course. Similar

* These authors contributed equally to this work.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

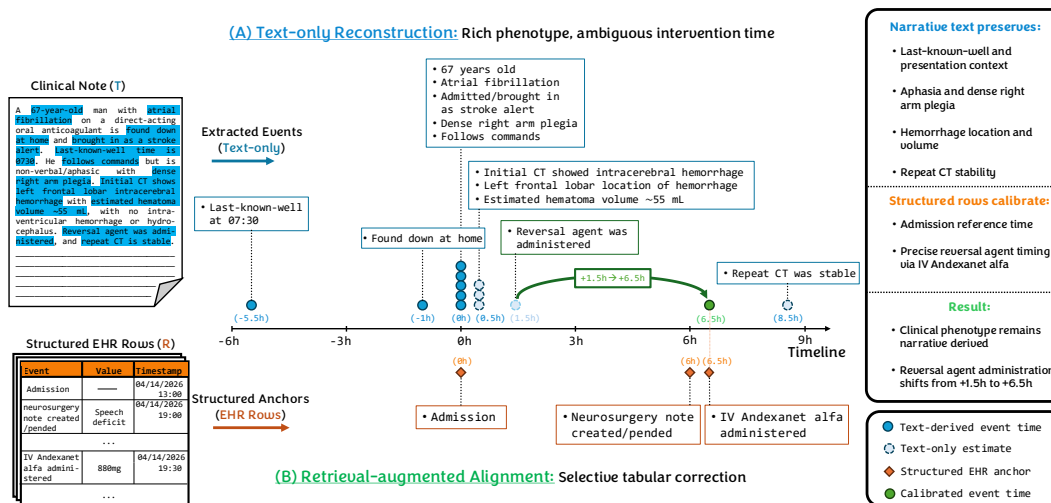


Fig. 1. Motivation for retrieval-augmented temporal grounding using a synthetic stroke-alert case constructed for illustration. **(A)** Text-only reconstruction can recover clinically meaningful narrative events, but intermediate event times may remain ambiguous. **(B)** Retrieved structured EHR rows provide partial temporal evidence that can calibrate text-derived events when clinically relevant, improving temporal placement while preserving narrative event content.

temporal challenges extend across critical illness: risk forecasts depend on when evidence becomes available; treatment-response analyses require distinguishing pre- and post-intervention measurements; and retrospective cohort definitions may change when the timing of symptom onset, diagnostic testing, or treatment initiation is misaligned. This increasing focus on temporal structure is evident in early-warning systems,³ studies showing that predictive utility varies with timing,⁴ and research on the temporal characteristics of sepsis subtypes and progression.⁵

The growing emphasis on time in critical care reveals an important methodological gap: electronic health records (EHRs) are not organized as temporally grounded patient trajectories. Critical-care repositories such as MIMIC-III/IV^{6,7} make this challenge explicit by combining structured time-series measurements with narrative clinical documentation. These modalities provide complementary but asymmetric evidence for timeline construction (Figure 1). Structured rows offer relatively precise temporal information from laboratory results, medications, procedures, vital signs, diagnoses, and administrative events. Clinical narratives, by contrast, provide richer accounts of the patient course, including symptoms, clinician assessments, patient-reported histories, functional status, disease progression, and contextual modifiers. For example, a note describing “increasing lethargy and peripheral mottling” may precede an increase in Sequential Organ Failure Assessment (SOFA) scores. Similarly, documented oliguria despite stable creatinine may indicate clinically meaningful states that structured measurements do not fully capture.

Recent studies point to both the promise and the difficulty of this problem. Multimodal annotation of discharge summaries aligned with structured EHR data demonstrated that access to both modalities can improve the precision of inpatient event timelines, highlighting the value of structured rows as temporal evidence when aligned with the narrative record.⁸ At the same time, text-only timeline reconstruction remains challenging in realistic EHR-style notes, where dense event structure, abbreviations, and semi-structured formatting make

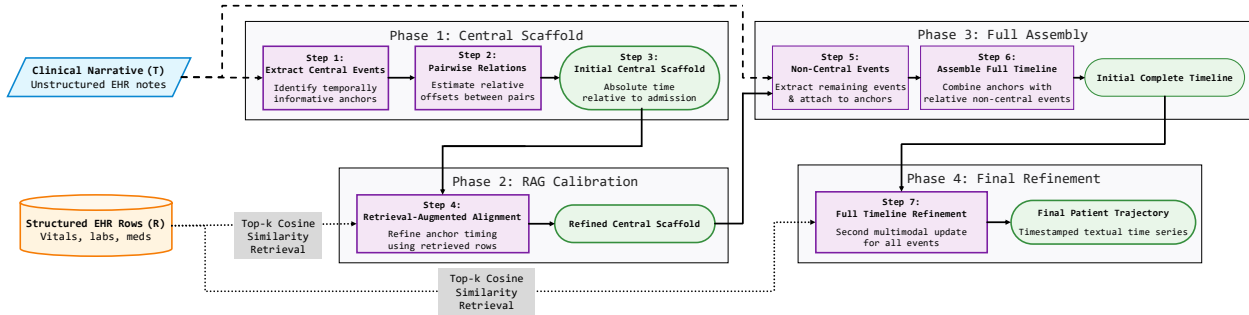


Fig. 2. Overview of the scaffolded retrieval-augmented timeline reconstruction workflow. The method first extracts central events from the clinical narrative, estimates pairwise temporal relations, and builds an initial central-event scaffold. Retrieved structured EHR rows are then used to calibrate this scaffold. Non-central events are extracted and attached relative to central events, after which a final retrieval-augmented refinement step calibrates the complete timeline.

both event recovery and time assignment difficult.^{9,10} Together, these findings motivate more explicit alignment between text-based and tabular approaches, with the goal of improving temporal precision while preserving the richer event content available in narrative text.^{11,12}

In this work, we investigate whether aligning retrieved structured EHR evidence with text-derived event timelines can improve the temporal precision of reconstructed patient trajectories. Our retrieval-augmented multimodal approach integrates structured rows directly into timeline reconstruction rather than treating textual and tabular data as separate views. The multistep pipeline first identifies temporally informative central events and uses them to create an initial scaffold of the patient trajectory. It then positions additional events relative to this scaffold and refines the resulting timeline using retrieved structured evidence. This decomposition reflects a clinically natural distinction between events that form the backbone of an encounter and those whose timing is best interpreted relative to that backbone. Using this framework, we assess whether and where multimodal alignment improves on text-only timeline reconstruction. Specifically, we examine the effects of introducing structured evidence at different stages of the reconstruction workflow, determine whether improvements are concentrated in temporal metrics rather than event recovery, and analyze which prognostically relevant information appears in text but is missing or delayed in tabular data.

Contributions. In summary, the paper makes the following contributions. **(i) Temporal grounding formulation:** We formulate clinical timeline reconstruction as retrieval-augmented temporal grounding of narrative-derived patient trajectories using structured EHR evidence. **(ii) Multimodal scaffolded reconstruction workflow:** We introduce a workflow that uses anchor events as a temporal backbone and retrieved structured rows as partial evidence for timestamp calibration. **(iii) Multimodal timeline evaluation:** We evaluate multimodal calibration on 40 discharge summaries with aligned structured EHR rows and manual gold-standard timelines, testing whether structured evidence improves temporal localization more than event recovery. **(iv) Information missing from tabular data:** We provide an empirical analysis of clinically relevant events that are absent, delayed, or semantically compressed in structured tabular data, motivating why unstructured text is necessary for constructing clinically meaningful patient trajectories.

2. Related Work

Clinical temporal information extraction. Clinical timeline reconstruction builds on a long tradition of temporal reasoning and temporal information extraction in medicine.^{13,14} Early annotation schemes and shared tasks, including i2b2 and THYME, focused on identifying clinical events, time expressions, and temporal relations such as before, after, overlap, and containment.^{15–18} Systems such as MedTime and cTAKES-temporal operationalized these ideas for clinical notes by extracting time expressions and temporal relations.^{19–21} Pre-LLM work also studied patient timelines and longitudinal record summarization, including timeline generation from clinical notes and HARVEST for summarizing longitudinal patient records.^{22,23} These studies established the importance of temporal structure in clinical narratives, but primarily focused on temporal annotation, relation extraction, or summarization rather than grounding narrative-derived absolute timelines in structured EHR evidence.

LLM-based clinical timeline reconstruction. Recent work has moved from relation-centered temporal extraction toward absolute event timelines. Leeuwenberg and Moens studied absolute event timeline extraction from English clinical reports, while Frattallone-Llado et al. showed that aligning discharge summaries with structured EHR data can improve inpatient event timeline precision.^{8,24} Textual time-series work has further shown that clinical narratives can be converted into event–time tuples for temporal forecasting, but these approaches are primarily text-only and do not use structured EHR rows as calibration evidence during reconstruction.^{9,10,25,26} Recent shared tasks such as ChemoTimelines have renewed interest in patient-level clinical timeline extraction using prompting, fine-tuning, and dictionary-enhanced LLM systems.^{27,28} Our work differs from these efforts by treating timeline reconstruction as multimodal temporal grounding: narrative text supplies event semantics, while structured EHR rows are retrieved as partial temporal evidence for timestamp calibration.

Asymmetric evidence in structured and narrative EHR data. Clinical machine learning studies have shown that structured and unstructured EHR data encode complementary information for phenotyping, prediction, and patient representation.^{11,12,29} Structured data provide dense timestamped measurements, but often omit symptoms, functional status, patient-reported histories, and clinical reasoning. Narrative notes preserve this context, but are often retrospective and not organized as chronological trajectories. This asymmetry makes clinical timeline reconstruction different from standard multimodal fusion: the goal is not to merge two complete views of the same event sequence, but to use structured rows as partial temporal evidence for a narrative-derived patient trajectory. Our retrieval-augmented scaffolded workflow is designed for this setting, selecting candidate structured rows for each text-derived event and updating timestamps only when the retrieved evidence is clinically relevant.

3. Methods

3.1. Clinical timeline task and dataset

Given a clinical narrative T and associated structured EHR rows R , the goal is to reconstruct a textual time series $\hat{S} = \{(e_i, \hat{t}_i)\}_{i=1}^N$, where each e_i is a standalone clinical event and each $\hat{t}_i \in \mathbb{R}$ is its predicted timestamp in hours relative to a case-specific reference time. When hospital admission is explicitly described, we use admission as $t = 0$; otherwise, we use the earliest

documented clinical encounter or presentation in the narrative. Events before the reference time receive negative timestamps, and events after it receive positive timestamps.

A clinical event is any patient-specific health-related mention that is semantically interpretable on its own and temporally locatable, including symptoms and signs, diagnoses, tests, procedures, treatments, medication administrations, major clinical states, outcomes, pertinent negatives, and termination or discontinuation events. Our event representation follows prior textual time-series work,^{9,25,26,30} but differs from short-span, concept-centric schemes such as i2b2.^{15,16} We represent events as standalone clinical findings: spans may include context needed for clinical specificity, conjunctive mentions are split when doing so improves clarity, and modifiers such as negation, uncertainty, severity, laterality, and intent are preserved when they affect event interpretation or temporal placement.

Our evaluation uses **40 discharge summaries** with manual gold-standard timelines and aligned structured EHR rows: **15 i2b2-derived discharge summaries** and **25 MIMIC-IV discharge summaries**. All cases have structured EHR rows aligned to the same hospital encounter. MIMIC-style critical-care records are well suited for this setting because they pair narrative clinical documentation with timestamped structured data, including laboratory values, medication administrations, vital signs, procedures, diagnoses, and administrative events.^{6,7} For each case, the narrative note provides the source text for event reconstruction, while the structured record provides external temporal evidence. We convert structured-row timestamps to hours relative to the same case-specific reference time used for the narrative timeline. Each structured row $r_m \in R$ is represented as $r_m = (\nu_m, x_m, s_m)$, where ν_m is the row name or clinical feature, x_m is the observed value when available, and s_m is the timestamp in hours relative to the reference time.

The 15 i2b2-derived cases were randomly sampled from eligible cases; the 25 MIMIC-IV cases were sampled at the patient level, using the first discharge summary for five randomly sampled patients and one randomly selected summary for 20 others, without enrichment for diagnosis, temporal richness, or encounter complexity. Three medical experts annotated disjoint subsets; every case was annotated once from the note and aligned structured record, so no adjudication or inter-annotator agreement was available (Supplementary Material §S1).

We treat structured rows as temporal evidence rather than gold-standard timeline events. Structured EHR tables contain dense repeated measurements, medication records, and machine-generated entries, many of which are not standalone narrative events. Conversely, clinically meaningful narrative events, such as symptoms, functional status, patient-reported histories, and clinical reasoning, may not have direct structured counterparts. Our method therefore uses structured rows to calibrate text-derived event times when clinically relevant, rather than forcing all structured rows into the reconstructed timeline.

3.2. *Scaffolded retrieval-augmented multimodal timeline reconstruction*

Our workflow is designed around three principles. First, it anchors before expanding: the model identifies temporally informative anchor events before reconstructing the full event set. Second, it retrieves candidate structured evidence rather than forcing every narrative event to align to a structured row. Third, it uses preserve-by-default calibration: event times are updated only when retrieved structured evidence is clinically relevant and temporally plausible. The

workflow proceeds in seven clinically interpretable steps (Figure 2).^a

- (1) **Step 1: Extract anchor events.** The model first identifies temporally informative anchor events, also referred to as central events, that define the encounter backbone or serve as temporal referents for other events. Examples include admission, symptom onset, treatment initiation, clinical deterioration, procedures, transfers, and discharge.
- (2) **Step 2: Estimate timing among anchors.** The model estimates pairwise time differences (in hours) between selected anchor events.
- (3) **Step 3: Build an initial anchor timeline.** The pairwise relations between anchor events are resolved into a provisional timeline backbone in hours relative to the case reference time. When estimates conflict, the model uses higher-confidence relations and the source narrative to construct the initial scaffold.
- (4) **Step 4: Calibrate anchors using structured evidence.** For each anchor event, retrieved structured EHR rows are provided to the calibration model together with the narrative and provisional scaffold. The model updates an anchor time only when the structured row clinically supports the change; otherwise, the text-derived estimate is retained.
- (5) **Step 5: Extract non-anchor supporting events relative to anchors.** After anchor calibration, the model extracts remaining non-anchor events and attaches each event to a referent anchor with a relative time offset and confidence score. These events include symptoms, exam findings, pertinent negatives, severity descriptors, functional status, medication details, and contextual clinical information.
- (6) **Step 6: Assemble the full timeline.** The calibrated anchor times and relative supporting-event offsets are combined into an initial complete timeline, converting all events into absolute times in hours relative to the case reference time.
- (7) **Step 7: Refine the complete timeline using structured evidence.** Structured rows are retrieved again for all reconstructed events, not only anchors. The model performs a final conservative calibration pass, updating timestamps when retrieved rows provide clinically consistent temporal evidence and preserving text-derived times otherwise. The output is a temporally grounded textual time series containing clinically meaningful narrative events with calibrated timestamps when structured evidence is relevant.

Retrieval and calibration module. The same retrieval module is used in Step 4 and Step 7. Structured rows are represented by concatenating the row event name and value; timestamps are excluded from the embedding text and passed separately to the calibration model. Rows with the same event name and value are merged, with timestamps represented in hours relative to the case reference time. We embed event strings and structured-row representations using MedTE (mohammadkhodadad/MedTE-cl15-step-8000), mean-pool and L2-normalize the embeddings, and rank rows by cosine similarity. For each event, we retrieve the top $k = 5$ structured rows with no fixed similarity threshold. The calibration model receives each row as `name | value | time | similarity`. Calibration follows a preserve-by-default rule: cosine

^aStep-specific prompts and the full reproducibility pipeline are available in our GitHub repository: https://github.com/Sayantankumar/multimodal_textual_tabular_alignment

similarity ranks candidates, irrelevant rows are disregarded, the initial time is retained when no relevant evidence justifies a change, and narrative context resolves repeated timestamps or conflicting rows. The complete decision rule appears in Supplementary Material §S2. We also evaluated $k = 3$, $k = 5$, and $k = 10$, finding negligible performance differences but increased runtime for larger k .

This design treats retrieval as evidence selection rather than forced temporal alignment. Methods such as dynamic time warping assume that two sequences share an approximately comparable event universe and can be aligned through a monotone path. Clinical narratives and structured EHR rows often violate this assumption: many narrative events have no structured counterpart, while structured tables contain dense repeated measurements, medication records, and charted observations that are not standalone narrative events. Our workflow therefore allows structured evidence to calibrate event times when appropriate without requiring every narrative event to have a tabular match.

3.3. *Experimental setup*

Baselines and ablations. We compare the default scaffolded multimodal workflow against text-only, single-step, stage-specific, and forced-alignment baselines. **Text-only reconstruction** uses the same scaffolded workflow but removes both structured-EHR calibration steps, testing whether retrieved structured evidence improves temporal localization beyond narrative-only reconstruction. **Single-step multimodal reconstruction** directly generates a complete timeline from the narrative and retrieved structured rows without constructing an anchor-event scaffold, testing whether the staged anchor/non-anchor decomposition is useful beyond simply adding retrieved evidence to the prompt. To isolate where structured evidence contributes, we compare two stage-specific ablations: **anchor-only calibration**, which calibrates the anchor scaffold but omits final full-timeline calibration, and **final-only calibration**, which reconstructs the full timeline from text before applying structured-EHR calibration. The default workflow applies both anchor-stage and final-stage calibration. Finally, **dynamic time warping (DTW)** serves as a forced temporal-alignment baseline that attempts to align text-derived events with structured rows through sequence alignment rather than retrieval-based evidence selection followed by clinical calibration.

Evaluation metrics. We evaluate predicted timelines along two complementary dimensions: *event recovery* and *temporal localization*, adapted from prior textual time-series work.^{9,25} Because predicted and reference timelines may differ in length and event wording, we first align predicted events to manually annotated reference events using a recursive best-match procedure. This produces a one-to-one alignment between predicted and reference events while accommodating unequal timeline lengths. **Event match rate** is the proportion of reference clinical events successfully aligned to a predicted event. Temporal metrics are computed on the matched subset. **Temporal concordance** (c-index) measures whether the predicted timeline preserves the relative ordering of events in the manual reference. **Area Under the Log-Time Cumulative Distribution Function (AULTC)** summarizes absolute timestamp discrepancy on a log-transformed time scale, with higher values indicating smaller timestamp errors across clinically meaningful ranges. Because concordance and AULTC are computed only over matched events, they should be interpreted jointly with event match rate. We estimate un-

Table 1. Performance on 40 discharge summaries at event-matching threshold 0.1 for unimodal and multimodal timeline reconstruction. Values are point estimates with bootstrap 95% confidence intervals in brackets. **Bold** indicates the best point estimate within each metric–modality column.

Model	Event match rate		Concordance		AULTC	
	Unimodal	Multimodal	Unimodal	Multimodal	Unimodal	Multimodal
DeepSeek R1	0.39 [0.32, 0.47]	0.39 [0.32, 0.47]	0.75 [0.69, 0.79]	0.75 [0.70, 0.80]	0.80 [0.76, 0.82]	0.79 [0.77, 0.82]
KimiK2-Instruct	0.26 [0.21, 0.29]	0.25 [0.22, 0.30]	0.73 [0.69, 0.76]	0.74 [0.68, 0.81]	0.68 [0.61, 0.73]	0.68 [0.62, 0.73]
GPT-OSS-120B	0.32 [0.27, 0.37]	0.32 [0.28, 0.37]	0.74 [0.70, 0.77]	0.75 [0.71, 0.80]	0.76 [0.72, 0.80]	0.77 [0.74, 0.81]
DeepSeek V3.2	0.49 [0.43, 0.56]	0.49 [0.43, 0.56]	0.74 [0.70, 0.80]	0.77 [0.73, 0.81]	0.71 [0.64, 0.78]	0.77 [0.73, 0.81]
GLM5	0.33 [0.27, 0.39]	0.33 [0.27, 0.40]	0.80 [0.78, 0.83]	0.82 [0.75, 0.84]	0.78 [0.73, 0.82]	0.78 [0.73, 0.82]
Qwen3.5-397B-A17B	0.38 [0.28, 0.47]	0.40 [0.29, 0.51]	0.78 [0.72, 0.81]	0.79 [0.78, 0.82]	0.74 [0.68, 0.77]	0.77 [0.74, 0.78]

certainty using case-level bootstrap confidence intervals. Each bootstrap sample resamples discharge summaries with replacement and recomputes event matching and timeline metrics, preserving within-case dependence among events. We report 95% confidence intervals for the reconstruction metrics using 200 bootstrap resamples. For direct multimodal–unimodal comparisons, we additionally used 10,000 paired case-level resamples and report confidence intervals for the mean within-case differences in Supplementary Material §S3, Table S1. All implementation and reproducibility details are presented in Supplementary Material §S8.

4. Results

4.1. *Structured EHR evidence calibrates event timing while preserving event recovery*

Table 1 summarizes timeline reconstruction performance for text-only reconstruction and multimodal reconstruction with retrieved structured EHR evidence. Across models, multimodal calibration has little effect on event match rate, indicating that structured rows do not primarily expand the set of events recovered from narrative text. Instead, structured evidence primarily affects temporal localization, with effects that vary by model and metric. These results support the role of structured rows as temporal evidence for text-derived events, rather than as a substitute event source.

Mean paired case-level temporal differences generally favored multimodal calibration: 4 of 6 models showed higher concordance and 3 of 6 showed higher AULTC, while no model showed lower performance on either temporal metric. However, uncertainty was substantial: no paired confidence interval for the multimodal–unimodal concordance difference excluded zero, and only DeepSeek V3.2 showed an AULTC interval entirely above zero (Supplementary Material §S3, Table S1). Performance was generally higher on the i2b2-derived subset, but modality effects were source-dependent (Supplementary Material §S5, Tables S3–S4).

Because event match rate depends on the semantic matching threshold used to align predicted and reference events, we also vary the event-matching threshold and examine the relationship between event recovery and temporal quality. Figure 3 shows AULTC and concordance as functions of event match rate across thresholds. Across operating points, multimodal reconstruction generally occupies a similar event-recovery range as text-only reconstruction while changing temporal performance, especially AULTC, for several models. This supports the same broad conclusion as Table 1: retrieved structured rows primarily affect temporal calibration rather than event discovery.

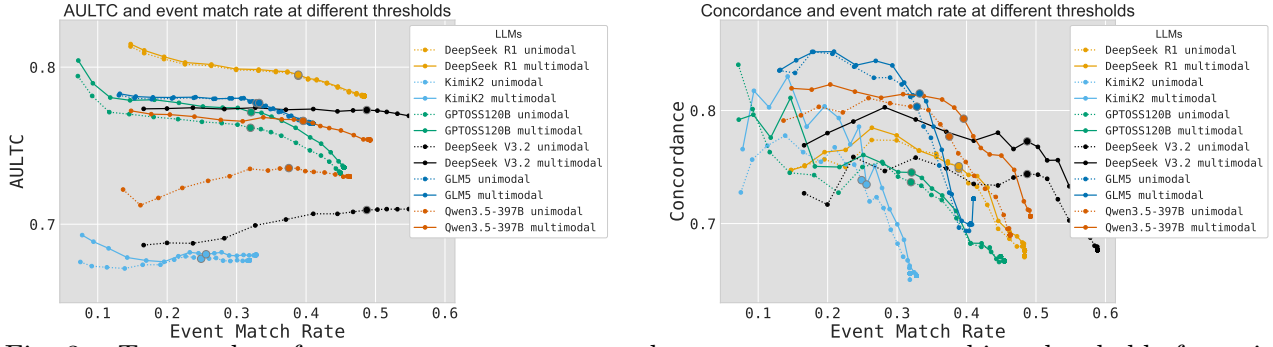


Fig. 3. Temporal performance versus event match rate across event-matching thresholds for unimodal and multimodal variants of each model on the 40-summary evaluation set. **(Left)** AULTC versus event match rate. **(Right)** Temporal concordance versus event match rate.

Scaffold ablations. Table 2A evaluates whether the staged scaffolded workflow is necessary, or whether similar performance can be obtained by generating a complete timeline in one step with retrieved structured rows. The single-step formulation performs worse than the full scaffolded variant, suggesting that clinical timeline reconstruction benefits from first identifying temporally informative anchor events before reconstructing the full event set. The stage at which structured evidence is introduced also matters. Final-stage calibration yields the highest event match rate, but the strongest temporal metrics are obtained when structured evidence is used both to calibrate the anchor scaffold and to refine the final assembled timeline. This supports the design choice of using structured rows first to stabilize the temporal backbone of the encounter and then to recalibrate the expanded trajectory.

Table 2. **(A)** DeepSeek V3.2 scaffold and calibration-stage ablations with bootstrap 95% confidence intervals. **(B)** Retrieval calibration compared with sentence-embedding and lexical Dynamic Time Warping (DTW) alignments.

A. Scaffold and calibration-stage ablations

Ablation	Event match rate		Concordance		AULTC	
	Unimodal	Multimodal	Unimodal	Multimodal	Unimodal	Multimodal
single-step	0.39 [0.35, 0.45]	0.39 [0.35, 0.44]	0.70 [0.65, 0.74]	0.71 [0.65, 0.75]	0.78 [0.74, 0.83]	0.77 [0.73, 0.83]
central only	—	0.46 [0.42, 0.51]	—	0.72 [0.66, 0.77]	—	0.75 [0.71, 0.80]
final only	0.51 [0.46, 0.56]	0.52 [0.46, 0.58]	0.68 [0.64, 0.72]	0.69 [0.64, 0.72]	0.72 [0.67, 0.78]	0.73 [0.69, 0.78]
central + final	0.49 [0.43, 0.56]	0.49 [0.43, 0.56]	0.74 [0.70, 0.80]	0.77 [0.73, 0.81]	0.71 [0.64, 0.78]	0.77 [0.73, 0.81]

Note: Central-only, final-only, and central + final correspond to updating the anchor scaffold only, the final timeline only, or both stages, respectively. The central-only condition yields a single final timeline and is undefined for the unimodal setting.

B. Retrieval calibration vs. explicit temporal alignment (Dynamic Time Warping)

Model	Event match rate			Concordance			AULTC		
	Retrieval	DTW(Sent)	DTW(Lex)	Retrieval	DTW(Sent)	DTW(Lex)	Retrieval	DTW(Sent)	DTW(Lex)
DeepSeek R1	0.39	0.39	0.39	0.75	0.75	0.74	0.79	0.80	0.79
KimiK2-Instruct	0.25	0.26	0.26	0.74	0.74	0.73	0.68	0.68	0.68
GPT-OSS-120B	0.32	0.32	0.32	0.75	0.73	0.73	0.77	0.77	0.77
DeepSeek V3.2	0.49	0.49	0.49	0.77	0.76	0.77	0.77	0.77	0.77
GLM5	0.33	0.33	0.33	0.82	0.81	0.81	0.78	0.78	0.77
Qwen3.5-397B-A17B	0.40	0.40	0.40	0.79	0.79	0.79	0.77	0.77	0.77

Explicit temporal alignment. We also compare retrieval-based calibration with dynamic time warping (DTW), an explicit temporal-alignment baseline that aligns text-derived events with structured rows as sequences (Table 2B). The two approaches achieve broadly comparable accuracy: DTW is slightly higher for some model-metric pairs, while retrieval-based calibration is slightly higher for others, and the differences are small relative to the variability reflected by the bootstrap confidence intervals in Table 1. Retrieval and DTW therefore show

Table 3. Structured-data gap analysis. Each text-derived event is compared with structured EHR counterparts from the same encounter.

Category	Definition	Count (%)
Well captured	Semantically adequate structured counterpart within the temporal window	2,827 (33.5%)
Complete absence	No structured counterpart identified	2,963 (35.1%)
Temporal mismatch	Semantically related structured row exists but occurs outside the temporal window	1,463 (17.3%)
Semantic distance	Retrieved row is not clinically equivalent to the narrative event	801 (9.5%)
Detail gap	Structured row omits clinically important attributes such as severity, laterality, duration, progression, or functional status	385 (4.6%)
Total textual events	—	8,439

comparable aggregate accuracy rather than quantitative superiority. Their principal distinction is structural: retrieval presents event-specific candidates and can retain a text-derived estimate when no adequate counterpart exists, whereas DTW imposes monotone sequence alignment across modalities with incomplete and asymmetric event coverage.

4.2. Complementary coverage of narrative and structured records

The main reconstruction results show that structured rows can improve event timing when used as calibration evidence. We next ask why structured records cannot simply replace narrative text and then characterize which narrative-derived events are poorly represented in structured EHR data; because this text-to-structure analysis does not assess structured-only events, it supports complementarity rather than completeness of either modality. We compare text-derived timelines against tabular records from the same encounters. The analysis uses the best-performing LLM-generated event timelines as the textual view, MIMIC structured data as the tabular view, and admission or the case reference time as the shared temporal anchor. For each text-derived event, we retrieve a candidate tabular counterpart using embedding-based similarity matching and evaluate it along two dimensions: semantic adequacy and temporal alignment. Semantic adequacy is assessed using retrieval-augmented scoring on a 0–1 scale, while temporal alignment is measured as the absolute difference in hours between the text-derived timestamp and the matched tabular timestamp. Based on these criteria, each event is classified as **well captured**, **complete absence**, **temporal mismatch**, **semantic distance**, or **detail gap**. Table 3 summarizes the distribution of these categories.

Across 8,439 text-derived events, only 33.5% are well captured by structured data, while 35.1% have no structured counterpart at all. When a structured counterpart exists, its timestamp is often useful: the median temporal discrepancy is 4.3 hours, with 58.4% of matched events within 6 hours, 69.5% within 12 hours, and 79.3% within 24 hours. The missing events are not merely low-value narrative detail; they include symptoms, functional status, temporal qualifiers, patient-reported information, severity descriptions, and clinical context. Representative complete-absence examples include **symptom descriptions** such as “substernal chest pain for 1 day” and “developed chest pain”; **denial statements** such as “no seizure-like activity” and “no deformity”; **temporal patterns** such as “pain resolved” and “improving clots on LE”; **severity descriptors** such as “EF 55%” and “non-small cell lung cancer stage IIIA”; and **causal or procedural context** such as “intubated for several days after second surgery” and “repair of leak from pouch.” These findings explain why structured rows are useful for temporal calibration but incomplete as representations of patient trajectories.

4.3. Qualitative case study: text supplies phenotype, tables calibrate time

Figure 1 illustrates the qualitative behavior of the pipeline using a synthetic stroke-alert case. The narrative supplies clinically decisive phenotype details, including lobar hemorrhage

location, hematoma volume, aphasia, plegia, command following, absence of intraventricular hemorrhage or hydrocephalus, and repeat-CT stability. The structured rows provide sparse but precise temporal anchors, including admission time and IV andexanet alfa administration. The text-only scaffold recovers the reversal event but places it shortly after the initial CT at approximately +1.5h; retrieval-augmented calibration links the event to the medication-administration row and shifts reversal to +6.5h relative to admission.

Pipeline output: calibrated event timeline

Event	Time (h)	Source / calibration note
Last known well at 0730	-5.5	Text clock time relative to admission
Found down at home	-1.0	Text-derived estimate
67 years old	0.0	Text, admission context
Atrial fibrillation	0.0	Text, admission context
Admitted/brought in as stroke alert	0.0	Table , anchored by admission row at 1300
Dense right arm plegia	0.0	Text, anchored to admission
Follows commands	0.0	Text, anchored to admission
Initial CT showed intracerebral hemorrhage	0.5	Text-derived estimate
Estimated hematoma volume ~55 mL	0.5	Text, anchored to initial CT
Left frontal lobar location of hemorrhage	0.5	Text, anchored to initial CT
Reversal agent was administered	6.5	Table , corrected by IV andexanet alfa at 1930
Repeat CT was stable	8.5	Text, inferred sequentially post-reversal

This case illustrates the asymmetry motivating the workflow. Text is necessary to recover clinically meaningful phenotype information, while structured rows calibrate selected time-sensitive events when relevant counterparts exist. The resulting trajectory could support human-reviewed cohort definition or eligibility logic by preserving narrative-only clinical details while grounding treatment timing in structured evidence. For example, the 2024 ENRICH trial comparing surgical evacuation versus medical management had eligibility criteria beyond what would be available in structured data alone.³¹ Structured data could show admission indication (hemorrhagic stroke), medication events (a confirmation proxy), and functional status scores (neurologic and disability level), but the notes contain key additional ENRICH eligibility features and timing: 55 mL hematoma volume ($30 < x < 80$), left frontal lobar location (lobar/anterior basal ganglionic), aphasia/plegia (new onset of neurological symptoms), no IVH/hydrocephalus and stable repeat imaging (feasibility), as well as localization of the eligibility timing window (within 24 hours of last known well time).

5. Discussion and Conclusion

Our results suggest that the primary benefit of multimodal alignment is temporal calibration rather than event discovery. Across models, structured evidence changed event recovery little, while mean paired case-level temporal differences generally favored multimodal calibration: 7 of 12 model-metric comparisons across concordance and AULTC were positive and none were negative. However, uncertainty in the magnitude of these improvements was substantial given the 40-case sample; paired confidence intervals excluded zero only for the DeepSeek V3.2 AULTC improvement. Source-stratified estimates also differed between i2b2-derived and MIMIC-IV cases (Supplementary Material §S3 and §S5). Within the text-primary formulation, narrative text defines the candidate event set and structured rows provide partial temporal evidence; the results therefore establish the value of narrative content for events represented in the note, not that narrative is the primary source of all clinically meaningful events.

The structured-data gap analysis explains why this distinction matters. Structured EHR data is not merely a noisier version of the narrative record; it offers a different view of the patient trajectory. When relevant counterparts exist, their timestamps can anchor timeline calibration. However, many clinically meaningful narrative events are absent, delayed, or semantically compressed in tabular form, including symptoms, functional status, temporal qualifiers, patient-reported information, and clinical context. These findings characterize asymmetric coverage within the present text-primary analysis; they do not assess structured-only events or establish that either source is complete.

The ablation results also support the scaffolded design. One-shot timeline generation requires a model to identify clinical events, infer relative timings, resolve vague temporal expressions, and decide how to use structured evidence in a single step. By first constructing an anchor-event scaffold, the workflow separates the temporal backbone of the encounter from supporting narrative details. This mirrors clinical reasoning, where events such as admission, diagnostic testing, treatment initiation, deterioration, procedures, and discharge provide temporal reference points for interpreting the rest of the patient course. Staged calibration then allows structured evidence to stabilize the backbone before refining the full trajectory.

Limitations: This study has several limitations. First, event match rates remain modest, so concordance and AULTC measure temporal quality conditional on successful event recovery rather than accuracy over the complete trajectory. Matched events were clinically heterogeneous, but unmatched events more often contained later-course and granular numerical documentation (Supplementary Material §S4, Table S2). Second, the 40-case sample provides only high-uncertainty evidence about stability across data sources and clinical settings; moreover, each case was annotated once by one of three medically trained annotators, without duplicate annotation or inter-annotator agreement. Third, the framework emits one continuous timestamp per event; when timing is vague or absent, it infers a point estimate from context, and the model-generated certainty flag is uncalibrated (Supplementary Material §S6, Table S5). Fourth, the text-primary design may omit clinically meaningful structured-only events. Structured timestamps also have source-specific semantics and may represent collection, ordering, administration, result availability, device recording, or documentation rather than clinical onset. Finally, the seven-call workflow has substantially higher inference cost than the single-step baseline.

Practical use and clinical impact: This framework is not intended for autonomous clinical decision-making, but for temporally grounded research and clinician-reviewed workflows. Reconstructed trajectories could support retrospective cohort construction, treatment-window definition, temporal phenotyping, leakage auditing, and downstream trajectory modeling. Future work should evaluate larger, independently annotated cohorts; support interval-valued timestamps, qualitative relations, and explicit time-unknown outputs; develop bidirectional event discovery that preserves source provenance and timestamp semantics; and test stage-specific cost-quality tradeoffs. Overall, the findings support a retrieval-augmented view in which narrative text supplies candidate-event semantics, structured rows provide partial and source-dependent temporal evidence, and preserve-by-default calibration can improve temporal grounding for some events when relevant evidence is available.

Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH) and utilized the computational resources of the NIH HPC Biowulf cluster. The contributions of the NIH author(s) are considered Works of the United States Government. S.N. was supported in part by the Carnegie Mellon University TCS Presidential Fellowship, and the Natural Sciences and Engineering Research Council of Canada (NSERC) Canada Graduate Research Scholarship — Doctoral (CGRS D) Fellowship. S.N. was also supported in part by an appointment to the National Library of Medicine Research Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the National Library of Medicine, National Institutes of Health. ORISE is managed by ORAU under DOE contract number DE-SC0014664. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH, NLM, DOE, ORAU/ORISE or the U.S. Department of Health and Human Services.

References

1. E. Kyriazopoulou, L. Liaskou-Antoniou, G. Adamis, A. Panagaki, N. Melachroinou, E. Drakou, K. Marousis, G. Chrysos, A. Spyrou, N. Alexiou *et al.*, Procalcitonin to reduce long-term infection-associated adverse events in sepsis. a randomized trial, *American Journal of Respiratory and Critical Care Medicine* **203**, 202 (2021).
2. C. W. Seymour, J. N. Kennedy, S. Wang, C.-C. H. Chang, C. F. Elliott, Z. Xu, S. Berry, G. Clermont, G. Cooper, H. Gomez *et al.*, Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis, *JAMA* **321**, 2003 (2019).
3. K. E. Henry, R. Adams, C. Parent, H. Soleimani, A. Sridharan, L. Johnson, D. N. Hager, S. E. Cosgrove, A. Markowski, E. Y. Klein *et al.*, Factors driving provider adoption of the trews machine learning-based early warning system and its effects on sepsis treatment timing, *Nature Medicine* **28**, 1447 (2022).
4. F. Kamran, D. Tjandra, A. Heiler, J. Virzi, K. Singh, J. E. King, T. S. Valley and J. Wiens, Evaluation of sepsis prediction models before onset of treatment, *NEJM AI* **1** (2024).
5. S. Noroozizadeh, J. C. Weiss and G. H. Chen, Temporal supervised contrastive learning for modeling patient risk progression, *Proceedings of Machine Learning Research* **225**, 403 (2023).
6. A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi and R. G. Mark, MIMIC-III, a freely accessible critical care database, *Scientific data* **3**, 1 (2016).
7. A. Johnson, T. Pollard, S. Horng, L. A. Celi and R. Mark, MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2) (2023).
8. G. Frattallone-Llado, J. Kim, C. Cheng, D. Salazar, S. Edakalavan and J. C. Weiss, Using multi-modal data to improve precision of inpatient event timelines, *Advances in Knowledge Discovery and Data Mining* **14648**, 322 (2024).
9. S. Noroozizadeh and J. C. Weiss, Reconstructing sepsis trajectories from clinical case reports using llms: The textual time series corpus for sepsis, *Proceedings of Machine Learning Research* **333**, 10 (2026).
10. S. Noroozizadeh, S. Kumar and J. C. Weiss, Forecasting clinical risk from textual time series: Structuring narratives for temporal ai in healthcare, *Proceedings of the AAAI Conference on Artificial Intelligence* **40**, 39080 (2026).
11. A. Moldwin, D. Demner-Fushman and T. R. Goodwin, Empirical findings on the role of struc-

- tured data, unstructured data, and their combination for automatic clinical phenotyping, *AMIA Joint Summits on Translational Science Proceedings* **2021**, 445 (2021).
12. T. M. Seinen, J. A. Kors, E. M. van Mulligen and P. R. Rijnbeek, Using structured codes and free-text notes to measure information complementarity in electronic health records: Feasibility and validation study, *Journal of Medical Internet Research* **27**, p. e66910 (2025).
 13. I. S. Kohane, Temporal reasoning in medical expert systems, in *MEDINFO 86: Proceedings of the Fifth World Congress on Medical Informatics*, eds. R. Salamon, B. Blum and M. Jørgensen (North-Holland, Amsterdam, 1986).
 14. L. Zhou and G. Hripcsak, Temporal reasoning with medical data—a review with emphasis on medical natural language processing, *Journal of Biomedical Informatics* **40**, 183 (Apr 2007).
 15. Ö. Uzuner, B. R. South, S. Shen and S. L. DuVall, 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text, *Journal of the American Medical Informatics Association* **18**, 552 (2011).
 16. W. Sun, A. Rumshisky and O. Uzuner, Evaluating temporal relations in clinical text: 2012 i2b2 challenge, *Journal of the American Medical Informatics Association* **20**, 806 (2013).
 17. W. Sun, A. Rumshisky and O. Uzuner, Annotating temporal information in clinical narratives, *Journal of biomedical informatics* **46**, S5 (2013).
 18. I. Styler, William F., S. Bethard, S. Finan, M. Palmer, S. Pradhan, P. C. de Groen, B. Erickson, T. Miller, C. Lin, G. Savova and J. Pustejovsky, Temporal annotation in the clinical domain, *Transactions of the Association for Computational Linguistics* **2**, 143 (2014).
 19. Y.-K. Lin, H. Chen and R. A. Brown, Medtime: A temporal information extraction system for clinical narratives, *Journal of biomedical informatics* **46**, S20 (2013).
 20. G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler and C. G. Chute, Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications, *Journal of the American Medical Informatics Association* **17**, 507 (2010).
 21. C. Lin, D. Dligach, T. A. Miller, S. Bethard and G. K. Savova, Multilayered temporal modeling for the clinical domain, *Journal of the American Medical Informatics Association* **23**, 387 (2016).
 22. A. Nikfarjam, E. Emadzadeh and G. Gonzalez, Towards generating a patient’s timeline: extracting temporal relationships from clinical notes, *Journal of biomedical informatics* **46**, S40 (2013).
 23. J. S. Hirsch, J. S. Tanenbaum, S. Lipsky Gorman, C. Liu, E. Schmitz, D. Hashorva, A. Ervits, D. Vawdrey, M. Sturm and N. Elhadad, HARVEST, a longitudinal patient record summarizer, *Journal of the American Medical Informatics Association* **22**, 263 (2015).
 24. A. Leeuwenberg and M.-F. Moens, Towards extracting absolute event timelines from english clinical reports, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **28**, 2710 (2020).
 25. J. Wang and J. C. Weiss, A large-language model framework for relative timeline extraction from pubmed case reports (2025).
 26. S. Kumar and J. C. Weiss, Temporally phenotyping glp-1ra case reports with large language models: A textual time series corpus and risk modeling, *medRxiv*, 2026 (2027).
 27. J. Yao, H. Hochheiser, W. Yoon, E. T. Goldner and G. K. Savova, Overview of the 2025 shared task on chemotherapy treatment timeline extraction, in *Proceedings of the 7th Clinical Natural Language Processing Workshop*, eds. A. Ben Abacha, S. Bethard, D. Bitterman, T. Naumann and K. Roberts (Association for Computational Linguistics, Virtual, oct 2025).
 28. T. M. Zhang, Z. Sun, S. Zeng, C. Li, N. F. Abernethy, B. D. Lam, F. Xia and M. Yetisgen, UW-BioNLP at ChemoTimelines 2025: Thinking, fine-tuning, and dictionary-enhanced LLM systems for chemotherapy timeline extraction, in *Proceedings of the 7th Clinical Natural Language Processing Workshop*, eds. A. Ben Abacha, S. Bethard, D. Bitterman, T. Naumann and K. Roberts

- (Association for Computational Linguistics, Virtual, oct 2025).
29. S. Liu, X. Wang and et al., Multimodal data matters: language model pre-training over structured and unstructured electronic health records, *IEEE Journal of Biomedical and Health Informatics* **27**, 504 (2022).
 30. S. Noroozizadeh, S. Kumar, G. H. Chen and J. C. Weiss, Pmoa-tts: Introducing the pubmed open access textual times series corpus (2026).
 31. G. Pradilla, J. J. Ratcliff, A. J. Hall, B. R. Saville, J. W. Allen, G. Paulon, A. McGlothlin, R. J. Lewis, M. Fitzgerald, A. F. Caveney *et al.*, Trial of early minimally invasive removal of intracerebral hemorrhage, *New England Journal of Medicine* **390**, 1277 (2024).

Do AI chatbots find what experts would? Effects of model, user role, and sample size on study retrieval for medical questions

Qingfang Liu¹, Qiao Jin², Joe D. Menke³, Thorsten Kahnt¹, Zhiyong Lu^{2,†}

¹*National Institute on Drug Abuse Intramural Research Program, National Institutes of Health, Baltimore, MD, USA*

²*National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*

³*School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL, USA*

[†]*Corresponding author: zhiyong.lu@nih.gov*

Large language model (LLM) chatbots are increasingly used to answer clinical questions with citations to relevant clinical studies. Prior research has largely focused on citation fabrication, leaving a gap in evaluating the quality of the retrieved studies and the factors driving their selection, especially for newer models with stronger reasoning capabilities. In this study, we evaluated three recent, general-purpose LLM chatbots: Claude Sonnet 5, Gemini 3.1 Pro, and ChatGPT GPT-5.5. We prompted the models with clinical questions adapted from 20 review questions in Issues 6 and 7 of the 2026 Cochrane Database of Systematic Reviews, simulating patient, clinician, and evidence-synthesis researcher user roles. For each review question, we queried all three chatbots across the three user roles, with four independent repetitions per chatbot–role combination, yielding 720 responses in total (3 chatbots $\times$ 3 user roles $\times$ 4 repetitions $\times$ 20 review questions). Each chatbot was asked to support its answers with primary clinical citations, which we then benchmarked against the included and excluded study sets of the Cochrane reviews. Included studies were those judged by Cochrane reviewers to meet the review-specific eligibility criteria, whereas excluded studies did not meet one or more of these criteria but could still be relevant to the clinical question. On average, a single chatbot response retrieved $39.2\% \pm 29.8\%$ (mean $\pm$ SD across all 720 responses) of the included-study set and $5.0\% \pm 9.4\%$ of the excluded-study set. Recall of Cochrane included studies varied significantly by model and user role. ChatGPT achieved higher recall than Claude or Gemini ($63.1\% \pm 29.5\%$ vs. $37.0\% \pm 23.8\%$ vs. $17.3\% \pm 13.1\%$; blocked permutation test, $p = 2.0 \times 10^{-5}$). The researcher role yielded higher recall than the clinician or patient roles ($42.8\% \pm 30.8\%$ vs. $38.6\% \pm 28.9\%$ vs. $36.1\% \pm 29.3\%$; $p = 2.0 \times 10^{-5}$). Controlling for publication year, citations per year, and open-access status, sample size was the only significant predictor of retrieval, with each doubling of sample size associated with 50% higher odds of retrieval (odds ratio 1.50, 95% CI 1.24–1.81, $p = 2.34 \times 10^{-5}$). These findings suggest that while LLM chatbots can retrieve some studies identified by expert reviewers, their performance varies by model and user role, and they exhibit a bias toward clinical trials with larger sample sizes. Collected LLM responses, analysis code, and supplementary materials are available at <https://github.com/QingfangLiu/llm-evidence-retrieval-bias>.

Keywords: Large Language Models; Evidence Retrieval; Retrieval Bias; Systematic Reviews; Cochrane Reviews; Evidence Synthesis; Medical Question Answering

1. Introduction

An increasing number of individuals are consulting chatbots to address medical questions, including requests to retrieve evidence from existing literature to better understand medical conditions,^{1,2} encouraged in part by strong chatbot performance on medical knowledge benchmarks.^{3,4} Yet the ability of general-purpose chatbots to identify and cite relevant clinical studies remains less well characterized,^{5–8} especially in light of prior findings that large language models can fabricate or hallucinate citations.^{9–13} Meanwhile, systematic reviews synthesize primary studies selected by experts to produce higher-quality conclusions for medical questions.¹⁴ This raises an important question: how closely do the studies cited by chatbots align with those included in systematic reviews?

Some prior work has begun to evaluate how well general-purpose chatbots retrieve primary studies, generally by comparing chatbot-cited studies against the included-study lists of published systematic reviews. These evaluations generally report low recall, whereas precision and citation fabrication vary across systems and evaluation methods. Somer et al.⁵ evaluated six publicly accessible systems against the 14 studies in an obstetric meta-analysis; the best-performing system, Claude 3.7, identified five studies. Chelli et al.⁶ tested three LLMs across 11 systematic reviews of rotator-cuff interventions, reporting precision of 0–13.4% and hallucination rates of 28.6–91.4%. Gwon et al.⁷ found that ChatGPT and Bing AI identified only one and two benchmark randomized trials, respectively, out of 24 randomized trials. Sidhu et al.¹⁵ compared the authenticity, quality, and geographic provenance of over 1,200 references across nine contemporary chatbot configurations. Low et al.¹⁶ found that general-purpose LLMs produced very few relevant, evidence-based answers (2–10% of questions). However, many of these studies relied on models predating advanced reasoning and agentic systems, leaving it unclear how modern models behave in these contexts. Additionally, much prior work has focused on a single medical domain,^{5–7} limiting the generalizability of their conclusions.

Outside biomedicine, retrieval-benchmark work such as LitSearch¹⁷ showed that retrieval performance varies substantially with query specificity and query construction. PaperAsk evaluated GPT-4o, GPT-5, and Gemini-2.5-Flash across a range of scientific fields and found that citation retrieval failed in 48–98% of multi-reference queries.¹⁸ Gao et al. evaluated five LLMs for 40 randomly selected original articles and found that the models failed to retrieve correct reference data 47.8% of the time.¹⁹ Related work has probed adjacent failure modes rather than recall itself: whether generated citations actually support their associated claims,¹⁰ how often citations are outright fabricated or contain metadata errors,⁹ and how reference-generation performance varies by model, discipline, and publication recency across large literature-review corpora.²⁰ Other work asked chatbots to generate search strategies for bibliographic databases such as PubMed and Embase rather than to select studies directly, but this setup may not reflect how general users interact with chatbots.²¹

A further consideration is that different users (e.g., patients, clinicians, evidence synthesis researchers) may seek primary clinical studies for medical questions,^{22,23} and whether such variation affects evidence retrieval remains an open question. Task-aligned role-play has been shown to help with reasoning benchmarks,²⁴ while assigning the model a persona does not reliably improve accuracy and can sometimes reduce it.²⁵ Persona information more broadly

shifts model predictions in ways that do not always track genuine human variation.²⁶ Prompt architecture and framing can induce systematic, non-neutral bias even when no persona is involved,²⁷ and patient question framing alone has been shown to change a model’s medical conclusions even when the underlying evidence is held fixed.²⁸ However, no prior study has tested whether a user’s self-identified role changes which primary studies a chatbot retrieves.

In this study, we evaluated three contemporary general-purpose chatbot systems (GPT-5.5, Claude Sonnet 5, and Gemini 3.1 Pro, all equipped with the ability to perform web searches) on medical questions adapted from 20 Cochrane reviews. Prompts were framed from the perspective of a patient, clinician, or evidence-synthesis researcher, and each role–review combination was repeated four times. GPT-5.5 achieved the highest recall of Cochrane included-study sets, and evidence-synthesis researcher framing yielded higher recall than clinician or patient framing. After controlling for publication year, citation rate, and open-access status, larger sample size remained the only significant predictor of study retrieval. These findings indicate that chatbot-based evidence retrieval is incomplete, context-dependent, and biased toward larger clinical trials, with implications for patients who use chatbots for medical information, as well as for clinicians, LLM developers, medical AI researchers, and policymakers.^{29–31}

2. Methods

2.1. Review selection

We used Cochrane intervention reviews published in Issues 6 and 7 of the 2026 *Cochrane Database of Systematic Reviews*. These were the two most recent complete issues available at the time of the experiment and were published after the knowledge cutoff dates of all three chatbots (see **Table 1**). The 20 eligible intervention reviews covered diverse clinical areas, evaluating pharmacological or biologic interventions ($n = 7$), procedural, surgical, or laboratory techniques ($n = 4$), bedside-care or clinical-management strategies ($n = 3$), nutritional supplementation ($n = 2$), rehabilitative or conservative physical treatments ($n = 2$), behavioral interventions ($n = 1$), and telehealth or service-delivery interventions ($n = 1$). We excluded protocols and reviews addressing diagnostic, prognostic, or other non-intervention questions.

2.2. Experiment

We evaluated three widely used consumer chatbots using the reasoning-enabled model configurations in **Table 1**. We focused on the general-purpose chat interfaces rather than specialized Deep Research modes to evaluate evidence retrieval in ordinary question-answering settings rather than extended literature search. To characterize the broader capability of the evaluated model configurations, we report performance on Humanity’s Last Exam (HLE), a multimodal, frontier-level academic benchmark comprising 3,000 expert-authored questions across dozens of subjects and designed to probe advanced knowledge and reasoning. These scores show that the evaluated systems are broadly comparable frontier-level models, though HLE itself does not benchmark literature retrieval, the primary focus of the current work. Each response was generated independently in a fresh anonymous web session with no prior conversational context. Data were collected from the U.S. East Coast between mid- and late July 2026.

Table 1. Chatbot systems, configurations, and benchmark performance.

Model	Knowledge cutoff	Release date	Reasoning configuration	HLE (no tools)	HLE (with tools)
Anthropic Claude Sonnet 5	January 2026	June 30, 2026	Medium effort; thinking enabled	43.2%	57.4%
Google Gemini 3.1 Pro	January 31, 2025	February 19, 2026	Extended thinking	44.4%	51.4%
OpenAI GPT-5.5	December 1, 2025	April 23, 2026	High reasoning	41.4%	52.2%

Note: Knowledge cutoff is the latest date covered by a model’s documented training knowledge. Humanity’s Last Exam (HLE) is a benchmark of advanced knowledge and reasoning and is reported to provide context on general model capability; it is not a benchmark of evidence retrieval.

For each review question, we designed three prompts across distinct user roles, keeping wording identical except for the role stated at the beginning of each prompt (**Table 2**). Prompts explicitly instructed chatbots to rely on primary evidence (e.g., original research studies and clinical trials) rather than secondary evidence (e.g., systematic reviews, meta analyses, practice guidelines) when answering. Given that the chatbot systems had web-search access, this instruction was intended to avoid, or where that was not possible, reduce reliance on secondary sources, including the benchmark Cochrane reviews, when identifying studies to present in the response. Each prompt was run four times to capture response variability.

Table 2. Example of the three user-role prompt variants for one review question.

User role	Role-specific opening
Patient	I have high blood pressure, and I am considering taking medicine to help me lose weight.
Clinician	I am a clinician caring for people with high blood pressure who are considering taking medicine to help them lose weight.
Evidence-synthesis researcher	I am an evidence-synthesis researcher studying weight-loss medicines in people with high blood pressure.
Shared remainder of the prompt: I am trying to understand whether weight-loss medicines improve the health of people with high blood pressure. Find individual primary studies on this. Do not use systematic reviews, meta-analyses, narrative reviews, committee opinions, practice guidelines, or editorials or comments. List the primary studies at the end.	

2.3. Study identification

We used the full set of studies listed as “included” and “excluded” in the corresponding Cochrane review as the reference standard and analyzed the studies retrieved in each chatbot response. For each response, the unit of analysis was the set of uniquely identifiable primary studies referenced, rather than the number of citations or publications, consistent with the

distinction between studies and publications in standard Cochrane methodology.

Study identification was performed in two stages. First, we used an AI-assisted workflow to extract citation information and propose candidate matches. Second, one author manually verified all proposed matches against the Cochrane review’s study records and, where necessary, external bibliographic databases. Because the prompts did not specify a citation format, we considered a study identifiable when the information provided (e.g., a study or trial name, author and year, title, PMID or DOI, registry identifier, or the destination of a citation link) was sufficient to resolve the underlying study. A citation with minor metadata errors (such as a wrong volume number) was still counted as a match, provided the rest of the information clearly pointed to one specific study; we report the metadata errors below. Unresolved citations received no matched-study credit.

For each chatbot response, we also assessed compliance with the instruction to rely on primary evidence by recording whether the response referenced or discussed a Cochrane review. For each Cochrane-included study, we resolved PMIDs for linked reports using RIS metadata and PubMed searches, then obtained citation counts from Semantic Scholar. We used the highest available citation count among linked reports to calculate citations per year, dividing by years since the Cochrane-reported publication year (minimum one year). Open-access status was the Semantic Scholar indicator for the report providing that citation count.

3. Results

3.1. *Number and composition of retrieved studies*

Across 720 responses, the mean number of studies retrieved per response was 10.66. Of the studies retrieved in each response, a mean of 61.9% were Cochrane-included studies, compared with 11.4% classified as Cochrane-excluded and 26.7% as other. Both retrieval volume and study composition varied across chatbots (**Figure 1**). Claude retrieved a mean of 12.32 ± 6.24 studies per response, of which 52.3% were classified as included—the lowest included-study proportion among the three chatbots. Gemini retrieved the fewest studies per response (3.87 ± 1.66) but had the highest included-study proportion (69.0%). ChatGPT retrieved the most studies per response (15.80 ± 6.60), of which 64.6% were classified as included. In contrast to the chatbot-level differences, study composition was similar across user roles, although researcher-role responses retrieved more studies than patient- or clinician-role responses.

3.2. *Included-study recall and recall consistency*

Across 720 responses, the mean per-response recall of Cochrane-included studies was $39.2\% \pm 29.8\%$ (mean $\pm$ SD), and the proportion of excluded studies cited was $5.0\% \pm 9.4\%$. Differences in mean included-study recall by chatbot and user role were assessed using separate blocked permutation tests with 50,000 permutations each. Mean included-study recall differed significantly by chatbot: $37.0\% \pm 23.8\%$ for Claude Sonnet 5, $17.3\% \pm 13.1\%$ for Gemini 3.1 Pro, and $63.1\% \pm 29.5\%$ for ChatGPT 5.5 ($p = 2.0 \times 10^{-5}$). Recall also differed by user role: $36.1\% \pm 29.3\%$ for patient, $38.6\% \pm 28.9\%$ for clinician, and $42.8\% \pm 30.8\%$ for researcher ($p = 2.0 \times 10^{-5}$). Recall consistency, measured as mean pairwise Jaccard similarity of retrieved Cochrane-included study sets across replicate responses, was highest for ChatGPT 5.5 (0.80), compared with



Fig. 1. **Composition and volume of studies retrieved by Cochrane status.** Stacked bars show the mean within-response proportion of distinct retrieved studies classified as Cochrane included, Cochrane excluded, or other, with bar length encoding mean studies retrieved per response on one shared scale across panels.

Claude Sonnet 5 (0.61) and Gemini 3.1 Pro (0.56), and was similar across user roles (patient, 0.64; clinician, 0.65; researcher, 0.67). These results are summarized in **Figure 2**.

3.3. Study overlap by chatbot and user role

We examined which studies were cited across multiple chatbots or user-role conditions and which were cited exclusively within a single group. Cochrane included and excluded studies were analyzed separately (**Figure 3**). Of the 442 included studies, 328 (74.2%) were cited in at least one response. When responses were grouped by chatbot and pooled across user roles, Claude cited 236 studies (53.4%), Gemini cited 119 (26.9%), and ChatGPT cited 303 (68.6%). Among the 328 cited studies, 107 (32.6%) were cited by all three chatbots. Eighty-three studies were cited exclusively by ChatGPT, 21 exclusively by Claude, and one exclusively by Gemini; an additional 105 were cited by both Claude and ChatGPT but not by Gemini.

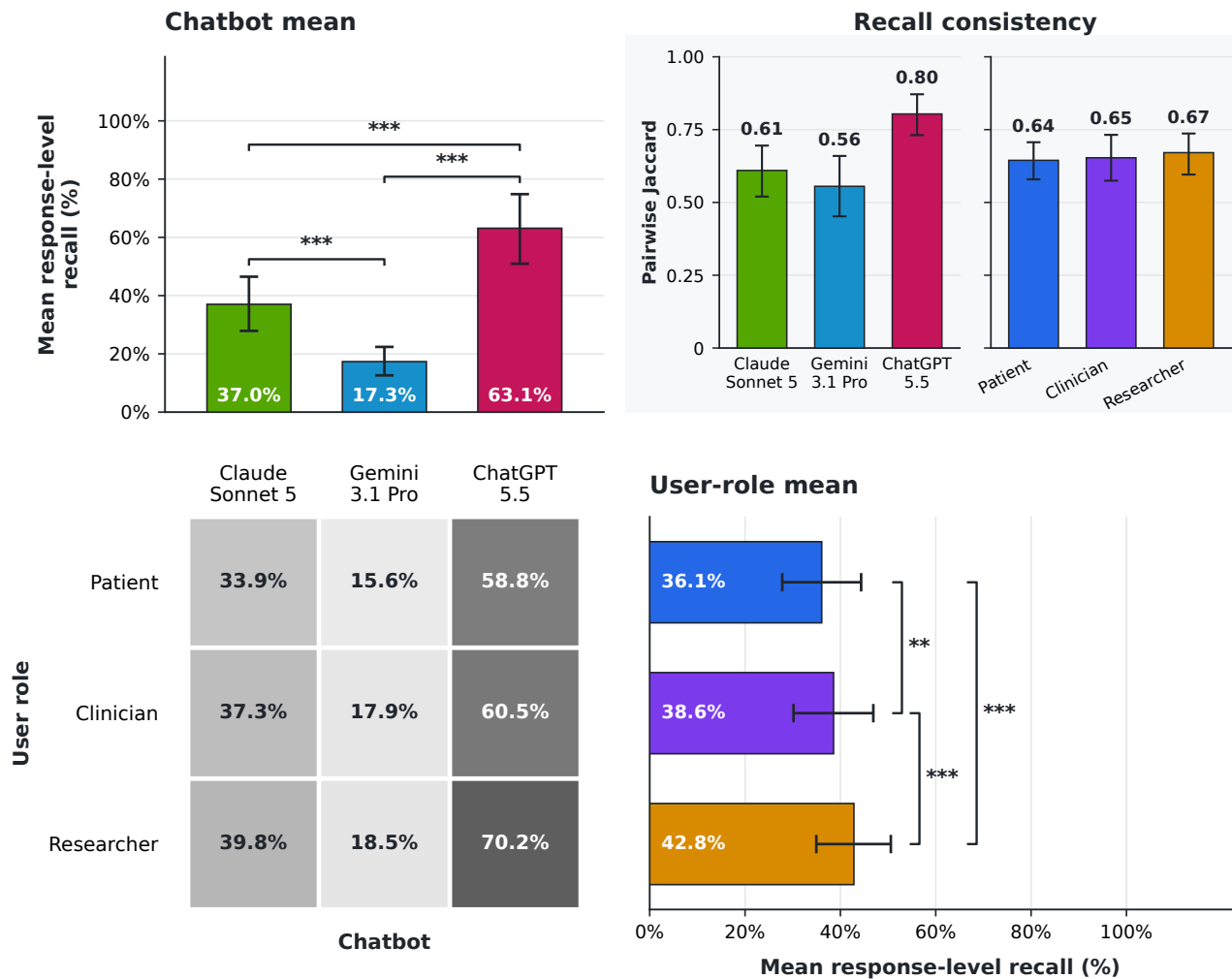


Fig. 2. **Cochrane included-study recall and recall consistency.** The heatmap shows mean response-level recall with marginal bars summarizing recall by chatbot and user role. The top-right panel shows recall consistency. Error bars indicate percentile 95% review-clustered bootstrap CIs. Pairwise differences were tested using blocked permutation tests (50,000 permutations), with Holm correction within each marginal dimension. Pairwise differences in mean recall were: Claude minus Gemini, 19.7 (14.3–25.7); ChatGPT minus Claude, 26.1 (19.3–33.5); ChatGPT minus Gemini, 45.8 (36.3–55.0); clinician minus patient, 2.5 (0.9–4.2); researcher minus clinician, 4.3 (2.5–6.2); and researcher minus patient, 6.8 (4.6–9.1). *** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$; ns, $P \geq 0.05$.

When responses were grouped by user role and pooled across chatbots, 249 studies (56.3%) were cited under the patient role, 272 (61.5%) under the clinician role, and 314 (71.0%) under the researcher role. Of the 328 cited studies, 234 (71.3%) were cited under all three roles, whereas 4, 8, and 43 were cited exclusively under the patient, clinician, and researcher roles, respectively. Thus, the included-study sets overlapped more strongly across user roles than across chatbots. Of the 932 excluded studies, 143 (15.3%) were cited in at least one response, and only 16 of the 143 cited excluded studies (11.2%) were cited by all three chatbots. Sixty-two of the 143 cited excluded studies (43.4%) were cited under all three roles. Thus, the excluded-study sets also overlapped more strongly across user roles than across chatbots.

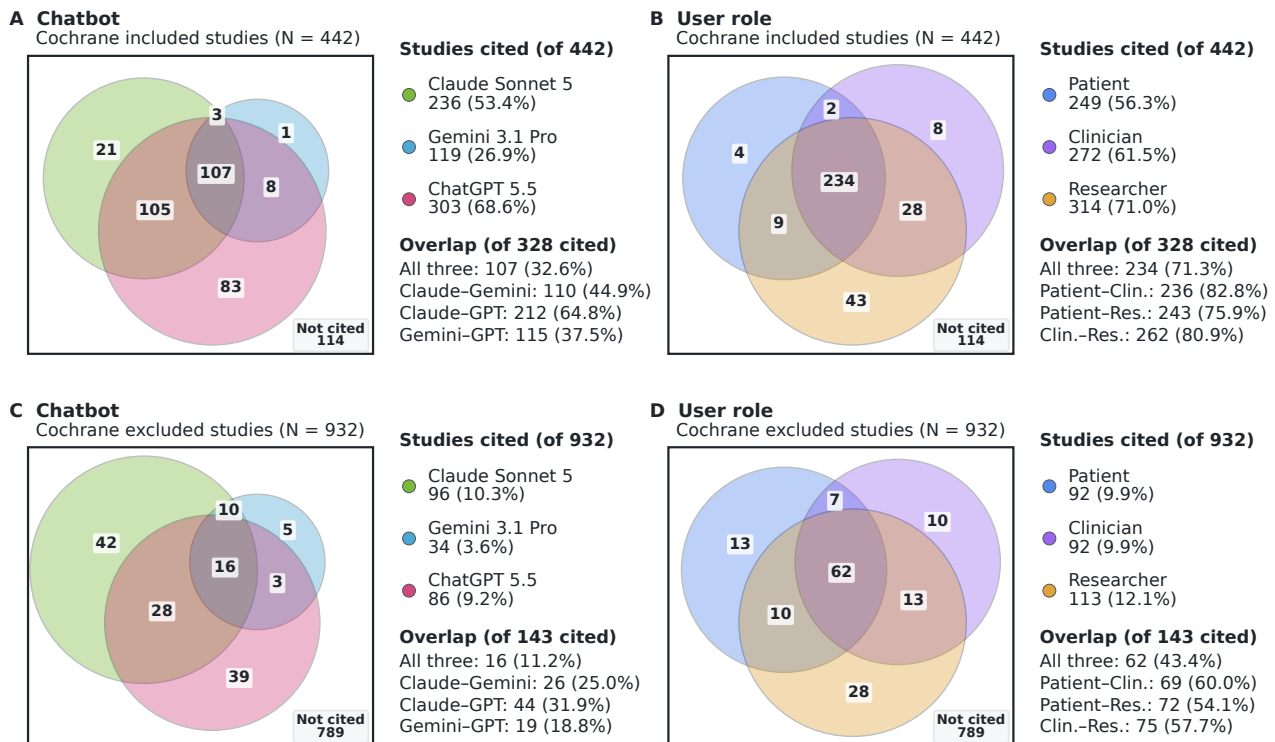


Fig. 3. **Overlap in Cochrane-included and Cochrane-excluded studies cited.** Venn diagrams show citations of 442 included and 932 excluded study labels across 20 Cochrane reviews. Circle areas are proportional within each panel to the number of studies cited by each chatbot or role. Region labels report the number of studies in each mutually exclusive shared or unique citation pattern.

3.4. Study characteristics associated with recall

Using the studies included in 20 Cochrane reviews as the reference standard, we compared recalled and unrecalled studies after pooling across all chatbots, user roles, and repetitions. A study was classified as recalled if it appeared in at least one response, and each Cochrane study label was counted once per review. Of 442 studies, 328 (74.2%) were recalled and 114 (25.8%) were not. Recalled studies were more recent (median publication year, 2014 vs. 2009; $p = 0.00058$), larger (median analyzed sample size, 195 vs. 97; $p < 0.001$), and more frequently cited (median citations per year, 5.50 vs. 2.33; $p < 0.001$). Open-access status did not differ significantly between recalled and unrecalled studies (56% vs. 45%; $p = 0.077$).

A multivariable logistic regression included publication year, log-transformed sample size, log-transformed citations per year, and open-access status (**Table 3**). Of 442 Cochrane-included studies, 330 (74.7%) had complete predictor data. Among the 330 studies, only sample size was significantly associated with recall, with each doubling of sample size associated with 50% higher odds of retrieval after natural-log transformation (adjusted OR, 1.50; 95% CI, 1.24–1.81; $p < 0.001$). Publication year, citations per year, and open-access status were not significantly associated with recall.

Table 3. Multivariable logistic regression of study recall.

Predictor	$\hat{\beta}$	Clust. SE	Clust. p	Naive p	OR	95% CI	VIF
Intercept	-34.9314	35.1336	0.320	0.194	—	—	—
Publication year	0.0164	0.0176	0.350	0.219	1.017	(0.98, 1.05)	1.30
log(sample size)	0.5880	0.1390	< 0.001***	< 0.001	1.800	(1.37, 2.36)	1.19
log(citations/year + 0.01)	0.2993	0.1574	0.057 [†]	0.010	1.349	(0.99, 1.84)	1.54
Open access	-0.1043	0.2598	0.688	0.740	0.901	(0.54, 1.50)	1.20

Note: SEs and primary p -values were clustered by review; naive p -values ignore clustering. OR = odds ratio; CI = confidence interval; VIF = variance inflation factor. McFadden pseudo- $R^2 = 0.1340$; log-likelihood = -146.64; likelihood-ratio test against the null model, $p < 0.001$. *** $p < 0.001$; [†] $p < 0.10$.

3.5. Characterizing Cochrane-excluded studies cited by chatbots

Given that the mean excluded-set retrieval was $5.0\% \pm 9.4\%$ per response, we characterized the retrieved excluded studies by classifying the 143 studies into 11 categories using the exclusion reasons reported by Cochrane review authors (**Table 4**). All 143 studies were assigned to at least one category; 29 studies (20.3%) had multiple exclusion reasons, so the categories were not mutually exclusive. The most common reasons for exclusion concerned study design, intervention, comparator, and population. This pattern is consistent with the structure of Cochrane eligibility criteria, which are typically defined according to the population, intervention, comparator, and eligible study designs specified for each review.³² Because our prompts were broader than the reviews' full eligibility criteria, some studies could be relevant to the chatbot questions yet ineligible for the more narrowly defined reviews.³³

3.6. Hallucination, fabrication and metadata errors

Hallucination is widely recognized as an issue in LLM-generated citations, but the term serves as an umbrella for several distinct phenomena:^{9,34,35} (1) metadata errors (e.g., incorrect lead author, year, journal, or page range) that can nonetheless be resolved to a real publication; (2) references that cannot be resolved to any real publication; and (3) misattribution of study content to an otherwise correct source. We report the first two categories here. Across reviews, 232 of 7,676 matched response–study citation rows (3.0%) contained metadata errors. Error rates varied substantially by model: 5.9% for Claude (175/2,956), 1.5% for ChatGPT (55/3,792), and 0.2% for Gemini (2/928). Differences by user role were smaller, with error rates of 2.7% for patients, 3.1% for clinicians, and 3.2% for researchers. In addition to metadata errors, we identified 23 citations that could not be resolved to any real publication; all occurred in Claude responses.

3.7. Secondary-evidence use and instruction adherence

We inspected the chatbots' responses for adherence to the instruction not to rely on secondary evidence, particularly Cochrane reviews. We coded each response along two dimensions: (1) *review identity*, indicating whether it referred to the corresponding 2026 Cochrane review,

Table 4. Exclusion categories among 143 Cochrane-excluded studies.

Category	n/143 (%)	Definition
Study design	37 (25.9)	Nonrandomized, observational, single-arm, pseudo-randomized, or otherwise ineligible designs
Intervention	34 (23.8)	Ineligible intervention identity, composition, route, dose, timing, or co-intervention
Comparator	28 (20.3)	Absent or ineligible control, or an ineligible comparison between intervention variants
Population	25 (17.5)	Ineligible participants or mixed populations without separable results for the eligible subgroup
Insufficient or unavailable data	15 (10.5)	Eligibility or eligible effects could not be established, separated, or obtained
Timing or follow-up	11 (7.7)	Ineligible treatment timing, study duration, or follow-up duration
Publication or trial status	8 (5.6)	Protocols, registrations without eligible results, unpublished, withdrawn, or retracted trials, or duplicate reports
Setting	6 (4.2)	Ineligible care or recruitment setting
Unspecified eligibility	5 (3.5)	Inclusion criteria were not met, but no specific criterion was identified
Outcome	2 (1.4)	Required outcomes were absent or ineligible
Unit of analysis	2 (1.4)	Ineligible allocation or analysis unit, or non-independent observations

an earlier or edition-unspecified Cochrane review, or no detected Cochrane review; and (2) *review role*, indicating whether the response explicitly stated that a review had been used to identify or verify primary studies or merely mentioned a review without indicating such a role. Cochrane reviews appeared in 66 Claude responses (27.5%). Of these, 17 responses explicitly stated that a review had been used to identify or verify primary studies, and 11 referred to the corresponding 2026 review. Only one response fell into both categories, referring to the corresponding 2026 review while also stating that a review had been used to identify or verify primary studies. We detected no references to Cochrane reviews in the Gemini or ChatGPT responses. Because we did not retain complete reasoning traces or external search histories, these results reflect only what was disclosed in the final responses. A chatbot may therefore have consulted a Cochrane review without mentioning it in its final answer.

4. Discussion

4.1. *Study findings*

Our findings show that chatbot retrieval was incomplete and selective. A single response retrieved an average of 39.2% of Cochrane-included studies, whereas pooling all responses identified 74.2% at least once. Recall varied more by chatbot than by user role. Larger sample size was the only significant predictor of retrieval after accounting for publication year, citation rate, and open-access status. Fully fabricated references were rare and occurred only in Claude responses, whereas citation metadata errors persisted across models and were more frequent for Claude than for ChatGPT and Gemini. Instruction adherence also varied by chatbot:

Claude sometimes referenced Cochrane reviews despite instructions not to rely on secondary evidence, whereas no such references were detected in ChatGPT or Gemini responses.

Our study has important implications. First, patients and clinicians should recognize that a chatbot response with citations reflects a model-specific and role-conditioned subset of the literature that may vary across repeated queries, rather than a comprehensive evidence summary. Second, systematic reviewers and evidence-synthesis researchers may find that general-purpose chatbots identify some relevant studies, but these tools should not replace reproducible database searches, formal eligibility screening, or citation verification.^{36,37} Third, AI developers working to automate evidence synthesis should make efforts to improve how systems interpret and apply population, intervention, comparator, outcome, and study-design (PICO-S) criteria. Accurate application of these criteria is essential for distinguishing the retrieval of high-quality evidence from the comprehensive retrieval of all relevant evidence.^{38,39}

4.2. Recall of primary studies

Recall in this study was generally higher than that reported in several earlier evaluations of general-purpose chatbots: mean per-response recall was 39.2% overall and 63.1% for ChatGPT, compared with 0%–13.7% in Chelli et al.,⁶ only one and two of 24 benchmark trials in Gwon et al.,⁷ and 35.7% for the best-performing model in Somer et al.⁵ This difference may reflect the use of newer LLMs with stronger reasoning and agentic capabilities, an explicit prompt requesting a list of primary studies, and a user role closely matched to the medical question.

We observed differences in recall across different chatbot models, and this appears to be largely driven by citation volume. ChatGPT cited a mean of 15.80 studies per response, approximately four times that of Gemini’s 3.87. The official release materials for all three models (GPT-5.5, Claude Sonnet 5, and Gemini 3.1 Pro) emphasize advanced reasoning and agentic capabilities, and their performance on Humanity’s Last Exam was broadly comparable. Yet, the models exhibited substantially different study-retrieval behaviors when answering medical questions. These discrepancies raise the question of whether they stem from differences in training data, retrieval architecture, or reasoning architecture, though the proprietary nature of these models makes this difficult to determine.

4.3. Factors affecting retrieval

Previous studies have reported that LLM-generated references tend to favor highly cited publications.^{20,40} In our study, however, after accounting for publication year, citations per year, and open-access status, trial size emerged as the clearest predictor of retrieval, with larger trials more likely to be retrieved. Citation rate was also positively associated with retrieval, but this association did not reach statistical significance. Several mechanisms may explain the preferential retrieval of larger studies. Search or retrieval-augmented generation pipelines may rank prominent “landmark” trials more highly, while larger trials may also be more strongly represented in model training data or other web-accessible sources. Chatbots may further select larger trials because sample size is interpreted as a signal of stronger or more reliable evidence. Alternatively, sample size may serve as a proxy for unmeasured characteristics such as journal visibility, trial registration, and broader scientific prominence. Because the search,

retrieval, and reranking processes of these consumer chatbot systems are proprietary and opaque, the current study cannot distinguish those possible mechanisms.

4.4. *Limitations*

Our study has several limitations. First, querying the models with questions from only 20 recent reviews limits the findings' breadth and generalizability, making this an exploratory test set rather than a comprehensive benchmark. Also, although Cochrane reviews provide a rigorous reference standard, their included-study sets are not immutable and may vary with review-specific judgments.⁴¹ Second, the study may not fully capture real-world patterns of use. Prompts were kept nearly identical across user roles to isolate the effect of this variable, but patients may express their evidence needs in more varied, less structured ways in practice.^{22,42} Models were also instructed to avoid secondary evidence to facilitate comparison with the Cochrane reference standard, but in real life, users are unlikely to add such constraints. Third, although we selected Cochrane reviews published after each model's knowledge cutoff to minimize potential training data exposure, the models may still have been exposed to earlier versions or other secondary sources. Moreover, despite being explicitly instructed to avoid using secondary evidence, the models differed in their adherence to this constraint.⁴³

4.5. *Future work*

We propose several directions for future work. First, future work should evaluate citations that fall outside the sets of included and excluded studies. Based on the present analysis, we cannot determine whether these citations represent irrelevant studies retrieved in error, newly published eligible studies, or other forms of related evidence.⁴⁴ Assessing their relevance will require independent adjudication and, in some cases, clinical expertise. Second, because our analysis focused on study retrieval, future research should also examine whether chatbot-generated conclusions accurately reflect the underlying evidence,^{10,12} appropriately balance benefits and harms, and communicate uncertainty. Such evaluations will likely require topic-specific expertise, particularly when the evidence is heterogeneous or clinically complex. Third, our analysis of cited study characteristics focused on publication year, trial size, citation rate, and open-access status. Additional factors such as the authors' country and abstract length warrant further study. Finally, future work should directly compare chatbot-based retrieval with conventional search-engine approaches, as well as newer agentic search systems, including Deep Research in ChatGPT and Gemini and domain-specific agentic RAG systems, to determine whether they can retrieve additional studies included in the reference Cochrane reviews that were missed by the chatbot systems evaluated in the current work.^{45–48}

5. *Acknowledgements*

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) are considered works of the US Government. Q.J. was also supported by the NIH Pathway to Independence Award K99LM014903. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the US Department of Health and Human Services.

References

1. B. Costa-Gomes, P. Tolmachev, E. Taysom, V. Sounderajah, H. Richardson, P. Schoenegger, X. Liu, M. M. Nour, S. Spielman, S. F. Way, Y. Shah, M. Bhaskar, H. Nori, C. Kelly, P. Hames, B. Gross, M. Suleyman and D. King, Public use of a generalist LLM chatbot for health queries, *Nature Health* **1**, 689 (2026).
2. R. B. Shpiner, Why patients ask the chatbot first, *Annals of Internal Medicine* (2026).
3. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Sementurs, A. Karthikesalingam and V. Natarajan, Large language models encode clinical knowledge, *Nature* **620**, 172 (2023).
4. Q. Jin, N. Wan, R. Leaman, S. Tian, Z. Wang, Y. Yang, Z. Wang, G. Xiong, P.-T. Lai, Q. Zhu, B. Hou, M. Sarfo-Gyamfi, G. Zhang, A. Gilson, B. Bhasuran, Z. He, A. Zhang, J. Sun, C. Weng, R. M. Summers, Q. Chen, Y. Peng and Z. Lu, Tutorial: Guidance on the use of large language models for medical research, *Nature Protocols* (2026).
5. S. Somer, M. Vardy, J. Eisenberger, M. Shapira, S. Mehler and M. Safrai, Evaluation of six large language models for study identification in an obstetric systematic review, *Minerva Obstetrics and Gynecology* **78** (2026).
6. M. Chelli, J. Descamps, V. Lavoué, C. Trojani, M. Azar, M. Deckert, J.-L. Raynier, G. Clowez, P. Boileau and C. Ruetsch-Chelli, Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis, *Journal of Medical Internet Research* **26**, p. e53164 (2024).
7. Y. N. Gwon, J. H. Kim, H. S. Chung, E. J. Jung, J. Chun, S. Lee and S. R. Shim, The use of generative AI for scientific literature searches for systematic reviews: ChatGPT and Microsoft Bing AI performance evaluation, *JMIR Medical Informatics* **12**, p. e51187 (2024).
8. A. Sarker, R. Zhang, Y. Wang, Y. Xiao, S. Das, D. Schutte, D. Oniani, Q. Xie and H. Xu, Natural language processing for digital health in the era of large language models, *Yearbook of Medical Informatics* **33**, 229 (2024).
9. W. H. Walters and E. I. Wilder, Fabrication and errors in the bibliographic citations generated by ChatGPT, *Scientific Reports* **13**, p. 14045 (2023).
10. K. Wu, E. Wu, K. Wei, A. Zhang, A. Casasola, T. Nguyen, S. Riantawan, P. Shi, D. Ho and J. Zou, An automated framework for assessing how well LLMs cite relevant medical references, *Nature Communications* **16**, p. 3615 (2025).
11. X. Wang, M. Tan, Q. Jin, G. Xiong, Y. Hu, A. Zhang, Z. Lu and M. Zhang, MedCite: Can language models generate verifiable text for medicine?, in *Findings of the Association for Computational Linguistics: ACL 2025*, eds. W. Che, J. Nabende, E. Shutova and M. T. Pilehvar (Association for Computational Linguistics, Vienna, Austria, July 2025).
12. Q. Jin, Y. Fang, L. He, Y. Yang, G. Xiong, Z. Wang, N. Wan, J. Chan, D. C. Comeau, R. Leaman *et al.*, Med-v1: Small language models for zero-shot and scalable biomedical evidence attribution, *arXiv preprint arXiv:2603.05308* (2026).
13. M. Topaz, N. Roguin, P. Gupta, Z. Zhang and L.-M. Peltonen, Fabricated citations: an audit across 2· 5 million biomedical papers, *The Lancet* **407**, 1779 (2026).
14. C. Lefebvre, J. Glanville, S. Briscoe, R. Featherstone, A. Littlewood, M.-I. Metzendorf, A. Noel-Storr, R. Paynter, T. Rader, J. Thomas and L. S. Wieland, Chapter 4: Searching for and selecting studies, in *Cochrane Handbook for Systematic Reviews of Interventions*, eds. J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page and V. A. Welch (Cochrane, 2025) Version 6.5.1; last updated March 2025.

15. R. S. Sidhu, A. Selvamogan and A. Boddy, Trust, truth and transparency: Analysing the references underpinning AI-generated surgical information, *The Annals of The Royal College of Surgeons of England* (2026).
16. Y. S. Low, M. L. Jackson, R. J. Hyde, R. E. Brown, N. M. Sanghavi, J. D. Baldwin, C. W. Pike, J. Muralidharan, G. Hui, N. Alexander, H. Hassan, R. V. Nene, M. Pike, C. J. Pokrzywa, S. Vedak, A. P. Yan, D.-h. Yao, A. R. Zipursky, C. Dinh, P. Ballentine, D. C. Derieg, V. Polony, R. N. Chawdry, J. Davies, B. B. Hyde, N. H. Shah and S. Gombor, Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems, *DIGITAL HEALTH* **11**, p. 20552076251348850 (2025).
17. A. Ajith, M. Xia, A. Chevalier, T. Goyal, D. Chen and T. Gao, LitSearch: A retrieval benchmark for scientific literature search, in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, (Association for Computational Linguistics, 2024).
18. Y. Wu, X. Liu, Y. Feng, J. Ding and X. Ma, Paperask: A benchmark for reliability evaluation of llms in paper search and reading (2025).
19. J. Gao, Y. Zhang, M. L. Disis and L. Zhang, Errors in ai-assisted retrieval of medical literature: A comparative study, *arXiv preprint arXiv:2603.22344* (2026).
20. X. Tang, X. Duan and Z. Cai, Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition, in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, (Association for Computational Linguistics, 2025).
21. W. S. W. Tam, N. Tung, S. X. Lee, G. Štiglic, T. Huynh and A. Tang, Development and evaluation of a generative AI chatbot for database searching in systematic review, *Journal of Nursing Scholarship* **58**, p. e70076 (2026).
22. K. Roberts and D. Demner-Fushman, Interactive use of online health resources: a comparison of consumer and professional questions, *Journal of the American Medical Informatics Association* **23**, 802 (2016).
23. N. Li, S. Orrange, R. L. Kravitz and R. A. Bell, Reasons for and predictors of patients' online health information seeking following a medical appointment, *Family practice* **31**, 550 (2014).
24. A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, X. Zhou, E. Wang and X. Dong, Better zero-shot reasoning with role-play prompting, in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, (Association for Computational Linguistics, 2024).
25. M. Zheng, J. Pei, L. Logeswaran, M. Lee and D. Jurgens, 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*, (Association for Computational Linguistics, 2024).
26. T. Hu and N. Collier, Quantifying the persona effect in LLM simulations, in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Association for Computational Linguistics, 2024).
27. M. Brucks and O. Toubia, Prompt architecture induces methodological artifacts in large language models, *PLOS One* **20**, p. e0319159 (2025).
28. H. S. Yun, G. Kapoor, M. Mackert, R. Kouzy, W. Xu, J. J. Li and B. C. Wallace, This treatment works, right? evaluating LLM sensitivity to patient question framing in medical QA (2026).
29. J. G. Meyer, R. J. Urbanowicz, P. C. Martin, K. O'Connor, R. Li, P.-C. Peng, T. J. Bright, N. Tatonetti, K. J. Won, G. Gonzalez-Hernandez *et al.*, Chatgpt and large language models in academia: opportunities and challenges, *BioData mining* **16**, p. 20 (2023).
30. D. Weissenbacher, M. Shabbir, I. M. Campbell, C. T. Berdahl and G. Gonzalez-Hernandez, Enhancing medical knowledge in large language models via supervised continued pretraining on clinical notes, *medRxiv*, 2026 (2026).

31. S. S. Sahoo, J. M. Plasek, H. Xu, Ö. Uzuner, T. Cohen, M. Yetisgen, H. Liu, S. Meystre and Y. Wang, Large language models for biomedicine: foundations, opportunities, challenges, and best practices, *Journal of the American Medical Informatics Association* **31**, 2114 (2024).
32. J. E. McKenzie, S. E. Brennan, R. E. Ryan, H. J. Thomson, R. V. Johnston and J. Thomas, Defining the criteria for including studies and how they will be grouped for the synthesis, *Cochrane handbook for systematic reviews of interventions*, 33 (2019).
33. T. Edinger and A. M. Cohen, A large-scale analysis of the reasons given for excluding articles that are retrieved by literature search during systematic review, *AMIA Annual Symposium Proceedings* **2013**, 379 (2013).
34. M. Dassen, R. Kotula, K. Murray, A. Yates, D. Lawrie, E. Kayi, J. Mayfield and K. Duh, Factum: Mechanistic detection of citation hallucination in long-form rag (2026).
35. M. Bhattacharyya, V. M. Miller, D. Bhattacharyya, L. E. Miller and V. Miller, High rates of fabricated and inaccurate references in chatgpt-generated medical content, *Cureus* **15** (2023).
36. J. Clark, B. L. Barton, L. Albarqouni, O. Byambasuren, T. Jowsey, J. W. Keogh, T. Liang, C. Moro, H. M. O'Neill and M. A. Jones, Generative artificial intelligence use in evidence synthesis: A systematic review, *Research Synthesis Methods* **16**, 601 (2025).
37. H. Chen, Z. Jiang, X. Liu, C. C. Xue, S. M. E. Yew, B. Sheng, Y.-F. Zheng, X. Wang, Y. Wu, S. Sivaprasad *et al.*, Can large language models fully automate or partially assist paper selection in systematic reviews?, *British Journal of Ophthalmology* **109**, 962 (2025).
38. Y. Li, S. Datta, M. Rastegar-Mojarad, K. Lee, H. Paek, J. Glasgow, C. Liston, L. He, X. Wang and Y. Xu, Enhancing systematic literature reviews with generative artificial intelligence: development, applications, and performance evaluation, *Journal of the American Medical Informatics Association* **32**, 616 (2025).
39. S. K. Vallamchetla, O. Abdelkader, A. Elnaggar, D. Ramadan, M. M. I. Shourav, I. B. Riaz and M. P. Lin, Do it faster with picos: Generative ai-assisted systematic review screening, *Journal of biomedical informatics* **168**, p. 104860 (2025).
40. A. Algaba, C. Mazijn, V. Holst, F. Tori, S. Wenmackers and V. Ginis, Large language models reflect human citation patterns with a heightened citation bias, in *Findings of the Association for Computational Linguistics: NAACL 2025*, (Association for Computational Linguistics, 2025).
41. D. G. Hamilton, J. E. McKenzie, P.-Y. Nguyen, M. L. Rethlefsen, S. McDonald, S. E. Brennan, F. M. Fidler, J. P. Higgins, R. Kanukula, S. Karunanathan *et al.*, Evaluation of the replicability of systematic reviews with meta-analyses of the effects of health interventions, *Research Synthesis Methods*, 1 (2026).
42. Q. Zeng, S. Kogan, N. Ash, R. A. Greenes and A. A. Boxwala, Characteristics of consumer terminology for health information retrieval, *Methods of information in medicine* **41**, 289 (2002).
43. Y. Jiang, Y. Wang, X. Zeng, W. Zhong, L. Li, F. Mi, L. Shang, X. Jiang, Q. Liu and W. Wang, FollowBench: A multi-level fine-grained constraints following benchmark for large language models, in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. L.-W. Ku, A. Martins and V. Srikumar (Association for Computational Linguistics, Bangkok, Thailand, August 2024).
44. P. Garner, S. Hopewell, J. Chandler, H. MacLehose, E. A. Akl, J. Beyene, S. Chang, R. Churchill, K. Dearness, G. Guyatt *et al.*, When and how to update systematic reviews: consensus and checklist, *bmj* **354** (2016).
45. Z. Wang, Z. Chen, Z. Yang, X. Wang, Q. Jin, Y. Peng, Z. Lu and J. Sun, Empowering biomedical evidence exploration and synthesis with deep knowledge graph research, *Nature Machine Intelligence* **8**, 1142 (2026).
46. Z. Wang, C.-H. Wei, J. Chan, R. Leaman, C.-P. Day, C. Wu, M. A. Knepper, A. S. Farias, J. Rincon-Torroella, H. Slika, B. Tyler, R. H.-T. Nguyen, A. Indurkar, M. Hébert, S. Tian, L. He, N. Naffakh, A. Aseem, N. Wan, E. Y. Chew, T. D. L. Keenan and Z. Lu, Deeper-med:

Advancing deep evidence-based research in medicine through agentic ai (2026).

47. Y. Xi, J. Lin, Y. Xiao, Z. Zhou, R. Shan, T. Gao, J. Zhu, W. Liu, Y. Yu and W. Zhang, A survey of large language model-based search agents, in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. M. Liakata, V. P. Moreira, J. Zhang and D. Jurgens (Association for Computational Linguistics, San Diego, California, United States, July 2026).
48. J. Deng, J. Huang, Z. H. Wong, H. Liang, Q. Xu, B. Cui and W. Zhang, Data-centric perspectives on agentic retrieval-augmented generation: A survey, in *Findings of the Association for Computational Linguistics: ACL 2026*, eds. M. Liakata, V. P. Moreira, J. Zhang and D. Jurgens (Association for Computational Linguistics, San Diego, California, United States, July 2026).

MedCaseReasoning: Evaluating and learning diagnostic reasoning from clinical case reports

Kevin Wu^{1,*} Eric Wu^{1,*} Rahul Thapa¹
 Kevin Wei² Angela Zhang³ Arvind Suresh³ Jacqueline J. Tao³
 Min Woo Sun¹ Alejandro Lozano¹ James Zou¹

¹Stanford University

²University of Southern California

³University of California, San Francisco

**Equal contribution. Correspondence: kewu93@gmail.com*

Doctors and patients alike increasingly use Large Language Models (LLMs) to diagnose clinical cases. However, unlike domains such as math or coding, where correctness can be objectively defined by the final answer, medical diagnosis requires both the outcome and the reasoning process to be accurate. Currently, widely used medical benchmarks like MedQA and MMLU assess only accuracy in the final answer, overlooking the quality and faithfulness of the clinical reasoning process. To address this limitation, we introduce *MedCaseReasoning*, the first open-access dataset for evaluating LLMs on their ability to align with diagnostic reasoning documented in clinician-written case reports. The dataset includes 13,989 diagnostic question-and-answer cases, each paired with detailed reasoning statements extracted by an LLM from open-access medical case reports. We evaluate state-of-the-art reasoning LLMs on *MedCaseReasoning* and find significant shortcomings in their diagnoses and reasoning: for instance, the top-performing open-source model, DeepSeek-R1, achieves only 48% pass@10 diagnostic accuracy and mentions only 64% of the reference reasoning statements (recall). However, we demonstrate that fine-tuning LLMs on the reasoning traces derived from *MedCaseReasoning* significantly improves diagnostic accuracy and clinical reasoning recall by an average relative gain of 41% and 29% (13.5 and 10.4 absolute percentage points), respectively. Our results showcase the importance of aligning LLMs to reason like human experts.

Keywords: Clinical reasoning; large language models; diagnostic accuracy; benchmark dataset.

1. Introduction

Large language models (LLMs) are increasingly deployed in medicine,¹ where they show promise in performing complex tasks like clinical reasoning² and disease diagnosis.³ While there are a number of benchmark datasets designed to assess the diagnostic capabilities of medical LLMs (e.g. MedQA,⁴ MMLU,⁵ MultiMedQA,⁶ and MedXpertQA⁷), these datasets all share the same limitation in that they only assess the correctness of the model's final answer. Unlike domains like mathematics⁸ or coding,⁹ where the reasoning process is secondary to the correct answer, medical diagnoses require both the final diagnosis and the thought process to be defensible. In clinical practice, physicians need to articulate coherent rationales, which is important for ensuring informed consent, meeting the standard of care, facilitating documentation and billing, and enabling clinical

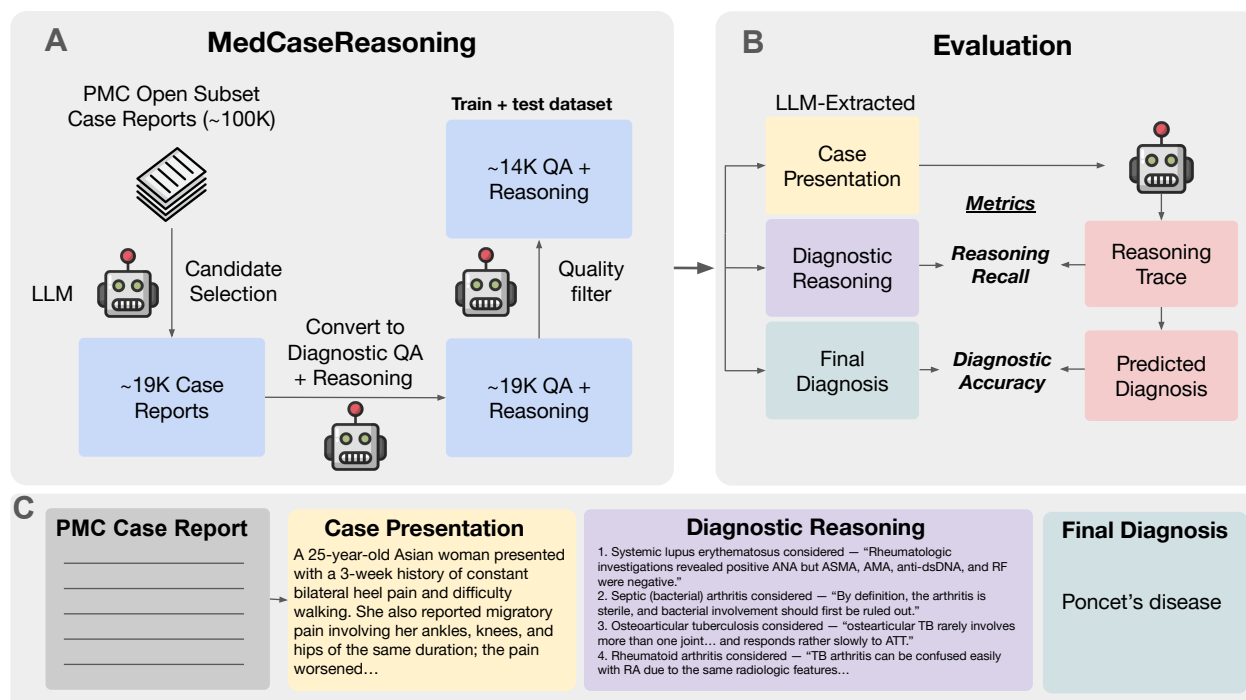


Fig. 1. A schematic of the MedCaseReasoning case processing pipeline. **A:** From the initial set of 98,994 PMC Open Subset Case Reports, we select 28,313 appropriate candidate cases, which are converted to QA format and then filtered for quality. The resulting 13,989 cases form the MedCaseReasoning dataset, with 897 test cases and 13,092 training cases. A full example is available in the Supplementary Material. **B:** For each test case, the case presentation is posed to an LLM to reason and then answer with a diagnosis. The reference reasoning extracted from the case report is compared with the LLM-generated reasoning to produce the reasoning recall score, and the case diagnosis is compared with the predicted diagnosis to produce diagnostic accuracy. **C:** The original case report text is converted into three sections: the case presentation that contains the relevant and sufficient patient information for making a diagnosis; the diagnostic reasoning that includes the diagnostic decision-making by the case author (in enumerated statements); and the final diagnosis.

collaboration.¹⁰ Mistakes or deficiencies in reasoning, even if the correct diagnosis is given, can risk case mismanagement. Highlighting this gap, a recent study found that even frontier models like GPT-4 could produce the correct diagnosis for incorrect reasons in up to a third of clinical scenarios.¹¹ Thus, evaluating and improving model capabilities to reason and answer correctly is crucial for LLMs to gain clinical viability and trustworthiness.

In this work, we propose **MedCaseReasoning**, a diagnostic reasoning dataset with two complementary goals: (1) an evaluation benchmark to assess how well LLM reasoning aligns with the diagnostic reasoning documented in clinician-written case reports, and (2) a training dataset for improving diagnostic reasoning in LLMs from reasoning traces. MedCaseReasoning is an open-access dataset of 13,989 diagnostic cases drawn from 800+ medical journals and 30+ specialties, created from publicly available case reports published in PubMedCentral.¹² To produce it, we develop a scalable, multi-step, clinician-validated pipeline that filters an initial set of 98,994 case reports and converts them into QA-format diagnostic cases, yielding a high-quality

test subset of 897 cases. First, we observe that `MedCaseReasoning` is a meaningfully difficult and unsaturated benchmark, with the top commercial model, *OpenAI o3*, achieving only 65% on pass@10 diagnostic accuracy. Second, we find that even top-performing reasoning models are substantially limited in their ability to reason like clinicians, where the best open-source model, *DeepSeek-R1*, only recalls 64% of the reasons provided in the case report. Third, we find that training on `MedCaseReasoning` can significantly improve the diagnostic and reasoning capabilities of LLMs: both diagnostic accuracy and reasoning recall increased after fine-tuning three top medical open-source LLMs, with an average relative gain of 41% and 29% (13.5 and 10.4 absolute percentage points), respectively. Importantly, after training on `MedCaseReasoning`, their performance on NEJM CPC,^{3,13} a held-out test dataset, also improved, showcasing the value of `MedCaseReasoning` in improving diagnostic generalizability.

Our work is the first medical benchmark dataset that explicitly evaluates the reasoning accompanying a diagnosis by comparing it to reasoning extracted from real-world case reports written by clinicians. These cases contain relevant differential diagnoses along with case presentations that present sufficient detail for an accurate diagnosis. Our findings underscore the significant challenges of aligning the reasoning of current top-performing LLMs with established clinical practice and provide a dataset for improving this alignment.

Throughout, *reference reasoning* denotes the reasoning statements an LLM extracts from a clinician-written case report; it is grounded in, but not identical to, the clinicians' own reasoning.

Related Works

Exam-style and literature-derived benchmarks. While commonly used medical benchmark datasets like MedQA,⁴ MMLU,⁵ and MedXpertQA⁷ include diagnostic reasoning questions, the clinical cases presented are typically shorter, textbook-style cases intended to evaluate students and test medical knowledge, not real-world clinical cases that physicians may encounter. These benchmarks are also multiple-choice and score only the selected answer. Several datasets construct evaluation items from real-world sources, such as PubMedQA,¹⁴ which generated QA examples based on PubMed articles, and MedAlign,¹⁵ which generated clinical vignettes from electronic health records. However, PubMedQA asks yes/no/maybe questions about research abstracts rather than posing open-ended diagnostic problems, and MedAlign targets clinician instruction-following over EHRs, which requires credentialed data access; neither provides reference diagnostic reasoning against which a model's reasoning can be scored. MedHELM¹⁶ broadens evaluation to a clinician-validated taxonomy of 121 medical tasks across 35 benchmarks, but it, too, scores model outputs rather than the reasoning that produced a diagnosis. By comparison, `MedCaseReasoning` is derived directly from clinician-authored case reports, which are valuable tools for advancing clinical practice,¹⁷ and pairs each open-ended diagnostic question with reference reasoning extracted from the report.

Real-world diagnostic cases. The New England Journal of Medicine Clinicopathologic Conferences (NEJM CPC) is a collection of case reports sourced from Massachusetts General Hospital in Boston, MA.³ Prior studies^{3,13,18} have leveraged NEJM CPC to evaluate model capabilities, but primarily focus on generating differential diagnosis (DDx) lists, not full reasoning traces. `MedCaseReasoning` improves upon this dataset in several ways. First, the collection of case reports from PubMedCentral represents a globally diverse set of clinician and patient backgrounds, whereas NEJM CPC represents the medical practices of physicians from a single

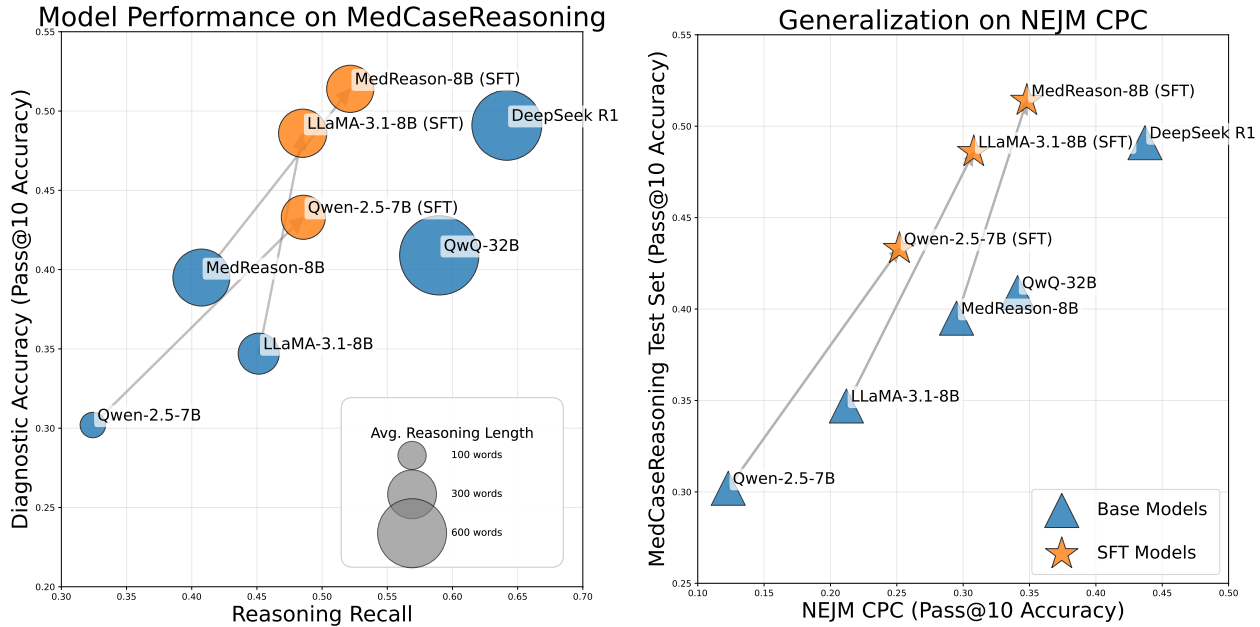


Fig. 2. (Left) Evaluation of LLMs on the MedCaseReasoning test dataset (N = 897), along with diagnostic accuracy and reasoning recall percentage. (Right) Model performance on the MedCaseReasoning test set and NEJM case studies (N = 302), showing a strong correlation between the two benchmarks (Pearson $r = 0.92$, $p < 0.001$, Pass@10 across the ten base and fine-tuned models). In both plots, accuracy is computed with pass@10 accuracy; circle size encodes the average length of the model-generated reasoning trace (trace length is correlated with reasoning recall; $r = 0.790$, $p = 0.0196$). LLaMA-3.1-8B and Qwen-2.5-7B are both Instruct variants (omitted for brevity).

hospital system on a specific population. Second, the NEJM CPC dataset consists of only 302 test cases, while MedCaseReasoning is comprised of 13,989 cases, allowing for more in-depth analysis as well as model fine-tuning. Third, MedCaseReasoning is derived from open-access articles on PubMedCentral, whereas NEJM CPC is available only under license.

Evaluating diagnostic reasoning. Several studies^{19,20} have employed grading rubrics applied to curated clinical vignettes, but are restricted to a handful of cases, not openly accessible, and focus on human-AI comparison rather than establishing benchmarks for standalone LLM reasoning performance. One study²¹ conducted manual clinician evaluations of reasoning traces from GPT-3.5 and GPT-4 to identify logical inconsistencies, but such methods are labor-intensive and inherently limited in scale, making them infeasible for large-scale benchmarking. MedCaseReasoning instead automates reasoning evaluation against reference reasoning for every case, with an LLM judge validated by a physician.

Training on medical reasoning traces. Recent models like HuaTuoGPT-o1²² and MedReason²³ involve supervised fine-tuning (SFT) on diagnostic reasoning traces, but the traces are distilled from larger proprietary models (e.g., GPT-4) rather than being grounded in reasoning authored by clinicians based on real patient cases. MedCaseReasoning provides training traces derived from clinician-written case reports, and we show that they improve models already trained on synthetic traces (Section 3).

Table 1. Comparison of clinical diagnosis benchmark datasets. `MedCaseReasoning` is the only and largest open-access dataset with clinician reasoning. Existing datasets do not contain expert reasoning.

Dataset	N	Open Access	Reasoning	Question Type	Question Source
MedQA ⁴	12,723	✓	✗	Clinical vignettes	USMLE
MedXpertQA ⁷	1,050	✓	✗	Clinical vignettes	Licensing exams
NEJM CPC ³	302	✗	✗	Real diagnostic cases	NEJM CPC
MedCaseReasoning	13,989	✓	✓	Real diagnostic cases	PMC Case Reports

2. Methods

2.1. Data curation

Case reports are typically written to highlight a rare or complex disease. A core tradeoff exists between the *novelty* and *usefulness* of a case report – models should learn from difficult cases, but should be able to practically deduce the diagnosis from the presented information. To this end, we produced a LLM-based pipeline with the goal of maximizing case novelty and usefulness. As expert verification on each step is prohibitively expensive, we perform our clinician validation only on the outputs of the last step.

- (1) **Data Provenance:** We started with the PMC Open Subset pulled from Jan 1st, 2005-April 27th, 2025. The query was filtered for free-access articles with article type "case report", which yielded 98,994 case reports. The article’s text was extracted from the main body in the XML file, along with the PMCID, publication date, title, journal name, and article link.
- (2) **Candidate Selection:** Next, we excluded case reports without the word *differential* (ie. a discussion on differential diagnoses), which reduced the set to 28,313 case reports. These initial candidate case reports were converted into diagnostic QA with *o4-mini*, a strong reasoning model that can be run at scale. When the dataset was built, *o4-mini* offered strong reasoning at a cost that made processing tens of thousands of articles feasible; the pipeline is model-agnostic and can be rerun with newer models. The full prompt is available in the Supplementary Material and instructs models to identify the core conflict within the case report and provide an information cutoff that would ensure the presented information does not give away the final diagnosis. The reasoning and diagnosis are extracted from the full report, so reference reasoning can occasionally cite findings after the cutoff (e.g., histopathology; see Section 5). In parallel, each candidate case report was scored according to the following criteria: 1) thoroughness of case presentation, 2) presence of explicit differential diagnosis, 3) dependence on integrative clinical reasoning, 4) transparency of diagnostic reasoning process, and 5) presence of stated final diagnosis. These criteria were developed by clinicians, and *o4-mini* was used to produce the actual scores. Criteria 1, 3, and 4 are given out of 1-5 points, whereas criteria 2 and 5 are yes/no responses. We filtered out cases where 1) thoroughness of case presentation was either 1 or 2 (seriously deficient

Table 2. An example of comparing reasoning traces against the case report reasoning. The reasoning from the case report is extracted using an LLM as an enumerated list. Then, each element of the list is referenced against the model’s reasoning trace, which gives us the “reasoning recall” for a given trace.

Reasoning	LLaMA-3.1-8B-Instruct (Base)	LLaMA-3.1-8B-Instruct (SFT on MedCaseReasoning)	Groundtruth (from MedCaseReasoning)
Trace	The patient has a history of Wilson disease, a genetic disorder that leads to copper accumulation in the body. However, the brain MRI shows no evidence of pathological copper deposition, and laboratory tests show normal copper metabolism – suggesting the current condition is not related to Wilson disease. The symptoms ... are consistent with a psychotic disorder. The history of Wilson disease and orthotopic liver transplantation may have led to post-liver-transplant psychosis ... Given the patient’s symptoms and history, post-liver-transplant psychosis is the most likely diagnosis.	We first considered acute graft rejection due to non-adherence with tacrolimus and mycophenolate, but this was ruled out because tacrolimus trough levels were low. Next, tacrolimus neurotoxicity was excluded for the same reason. We also considered copper overload from Wilson disease, but MRI showed no pathological copper deposition. Delirium was excluded; finally, we evaluated substance-induced psychosis but found “no other medical condition or substance use disorder.” With alternatives excluded, schizophrenia was suspected.	1. Wilson-disease recurrence — “no abnormalities, including pathological copper accumulation ... no evidence of abnormal copper metabolism ... psychosis due to recurrence of Wilson disease was ruled out.” 2. Tacrolimus-induced psychosis — “Tacrolimus-induced psychosis was ruled out because the trough tacrolimus concentration had not increased before onset of psychosis, and psychosis persisted after the trough value decreased.” 3. Substance-related or other medical causes — “No evidence of other medical conditions or a substance use disorder that could cause psychotic symptoms was found.”
Answer	Post-liver-transplant psychosis	Schizophrenia	Schizophrenia
Coverage	1/3 reasons recalled	3/3 reasons recalled	Groundtruth

or major gaps), and also removed cases that either did not discuss 2+ plausible alternatives or did not state a final diagnosis. The full prompt is available in the Supplementary Material. After filtering, there remained 19,428 total case reports.

- (3) **Quality Filter:** Next, to avoid model blind spots, we checked each of the generated case reports using a separate LLM (*gemini-2.5-pro*), which evaluated the generated case report’s faithfulness to the source article and the plausibility of each case report. The prompt is also available in the Supplementary Material. We removed cases that had any flags, leaving a final total of 13,989 cases. We created an initial testing subset of 897 cases where the transparency and integrative diagnostic reasoning scores were at least 4 or 5.

The case reports used in MedCaseReasoning span over 800 different medical journals, with

the top 10 shown in the Supplementary Material. Additionally, the diagnostic case prompts from `MedCaseReasoning` are significantly longer and more detailed than the ones from `MedQA` (see Supplementary Material). Furthermore, case report publication dates are heavily enriched towards years after 2020, with over 16% after Jan 1 2024 (see Supplementary Material). Additionally, our pipeline can be regularly updated, incorporating new case reports with low marginal cost.

To validate the case prompts, diagnostic reasoning, and final diagnosis extracted from each case report, a team of four board-certified physicians reviewed a total of 100 randomly selected cases, with each case reviewed by a single physician. The physician compared the extracted case prompt, reasoning, and diagnosis against the source article and answered three yes/no questions on hallucinations, faithfulness of the diagnosis, and faithfulness and relevance of the reasoning. The full questions can be found in the Supplementary Material.

2.2. Evaluating Diagnostic Accuracy in Models

We evaluate each model’s diagnostic correctness using LLM-as-a-judge, in accordance to previous literature (e.g.,^{3,23}). We adopt the same prompt used in prior work³ (see Supplementary Material), which has been validated to have high concordance with human raters, and use *gpt-4o-mini* as the judge model for its speed and accuracy.

Differential diagnoses typically contain between five to ten candidate diagnoses which are followed-up in a clinical setting. Also following the evaluation schema established by prior work,³ each model is evaluated a total of 10 times (temperature of 0.8 and top-p of 0.95), and the pass@N performance is recorded.

As an external validation to the `MedCaseReasoning` Test Set, we also evaluate models on case reports from NEJM Clinicopathologic Conferences. In particular, we use a subset of 302 cases from previous works^{3,13,18} as a gold-standard dataset of complex diagnostic cases. We manually extracted each case report with a personal license and fed the case presentation portion as the case prompt without any additional modifications. As NEJM CPC is not an open-access publication, we are not able to open-source this validation set. Additionally, no NEJM case reports were included in `MedCaseReasoning` or used in training any models.

2.3. Evaluating Reasoning Traces in Models

While some case studies include comprehensive reasoning steps, it is not guaranteed that case reports include an exhaustive set of reasoning steps. As such, we focus our study on evaluating how well models *recall* clinician-produced reasons (as opposed to evaluating *precision*). We define this metric as the “Reasoning Recall”. (Note: We were not able to evaluate reasoning recall for *OpenAI o3* and other frontier models where the full reasoning traces are not available via API call. While they can produce reasoning tokens when prompted (i.e. with <think> tags), the comprehensive reasoning that is used to inform a diagnosis is not accessible. We do not score such prompted explanations, which are produced separately from the hidden reasoning behind the diagnosis and would not be comparable to the native traces of open models.)

Definition 1 (Reasoning Recall). Let N be the total number of cases. For each case i , let

$$R_i = \{\text{groundtruth reasons from case report}\},$$

$$T_i = \{\text{reasons from model reasoning trace}\}.$$

For case i , the recall rate is

$$c_i = \frac{|R_i \cap T_i|}{|R_i|}.$$

The *Reasoning Recall* is the average of these rates:

$$\text{RR} = \frac{1}{N} \sum_{i=1}^N c_i = \frac{1}{N} \sum_{i=1}^N \frac{|R_i \cap T_i|}{|R_i|}.$$

For example, a reasoning recall of 50% indicates that on average, a model’s reasoning trace will cover half of the reasons provided by the case report. We evaluate the reasoning trace associated with its best-of-10 response, where we use the trace from a randomly selected correct response (or incorrect if there are no correct responses). The evaluation is implemented using an LLM-as-a-judge, where *o4-mini* is instructed to return a JSON with a decision on whether each of the groundtruth reasons are found in the reasoning trace. The prompt used is found in the Supplementary Material.

For example, in Table 2 the reference reasoning has three statements (Wilson-disease recurrence, tacrolimus-induced psychosis, and substance-related causes, each ruled out), so $|R_i| = 3$. The judge checks each statement against the full trace: the base *LLaMA-3.1-8B-Instruct* trace addresses only Wilson disease ($c_i = 1/3$), while the fine-tuned trace addresses all three ($c_i = 1$).

Validating the reasoning judge. We validate our LLM-as-a-judge for determining reasoning recall with verification from a board-certified physician. The physician was given $N=33$ cases and was asked to verify the LLM judge’s decisions on a total of 89 pairs of groundtruth reasons and model thinking traces. The cases were randomly sampled across all evaluated models. For example, for a given case report that contained three reasoning steps, the physician was asked to cross-check each step against the entire model reasoning trace to see if it was considered. The physician found 84/89 (94.4%, 95% CI of 87.5%-97.6%) of pairs of reasons and thinking traces to be correctly assessed by the model, and a total of 31/33 (93.9%, 95% CI of 80.4%-98.3%) cases to be judged correctly on every one of their reasoning statements (i.e., the physician agreed with all of the judge’s found/not-found decisions for that case).

2.4. Supervised Fine-Tuning on Case Report Reasoning

In addition to evaluating model reasoning traces, we explore whether fine-tuning on extracted reasons directly can improve model performance on diagnostic accuracy and reasoning recall. One technical challenge is that the extracted diagnostic reasoning is formatted as an enumerated list of summary points and quotations, rather than a cohesive reasoning trace. Naturally, we can use LLMs to "stitch" the list of points into reasoning traces without adding new information. Given the case prompt, reasoning points, and final diagnosis, the stitching model rewrites all points as one narrative, reusing quotations nearly verbatim and adding no reasoning of its own. We provide the prompt used to perform this step in the Supplementary Material. The stitching process may introduce biases if a stronger model introduces its own priors into the reasoning traces. To control for this, each training

Table 3. Performance of models on MedCaseReasoning’s test set (N=897). Each model is evaluated on reasoning recall and diagnostic accuracy (broken down into Pass@1, Pass@5, and Pass@10). Each cell reports the point estimate, with its bootstrapped 95% confidence interval (lower–upper bound) in subscript.

Reasoning recall is not available for *OpenAI o3* as the reasoning traces are not provided via API. We additionally perform supervised fine-tuning (SFT) on three open-source models on the training split of MedCaseReasoning and find significant improvements in both reasoning recall and diagnostic accuracy (average absolute gains of 10.4 and 13.5 percentage points in reasoning recall and Pass@10, respectively).

MedCaseReasoning Test Set				
Base Models	Reasoning Recall	Pass@1 Acc.	Pass@5 Acc.	Pass@10 Acc.
OpenAI o3	N/A	0.470 _{0.440-0.500}	0.609 _{0.579-0.639}	0.645 _{0.618-0.675}
DeepSeek R1	0.642 _{0.616-0.667}	0.320 _{0.291-0.349}	0.447 _{0.417-0.478}	0.480 _{0.450-0.510}
QwQ-32B	0.590 _{0.560-0.619}	0.272 _{0.245-0.302}	0.371 _{0.341-0.400}	0.398 _{0.368-0.428}
MedReason-8B	0.407 _{0.383-0.431}	0.248 _{0.224-0.275}	0.331 _{0.303-0.363}	0.382 _{0.353-0.412}
LLaMA-3.1-8B-Instruct	0.451 _{0.428-0.475}	0.161 _{0.138-0.184}	0.281 _{0.252-0.311}	0.332 _{0.304-0.360}
m1-7b-23k	0.495 _{0.440-0.551}	0.155 _{0.133-0.177}	0.238 _{0.211-0.264}	0.291 _{0.262-0.321}
Qwen-2.5-7B	0.324 _{0.301-0.347}	0.174 _{0.152-0.197}	0.252 _{0.223-0.279}	0.287 _{0.259-0.316}
Fine-Tuned Models	Reasoning Recall	Pass@1 Acc.	Pass@5 Acc.	Pass@10 Acc.
MedReason-8B (SFT)	0.522 _{0.498-0.545}	0.303 _{0.276-0.331}	0.430 _{0.400-0.459}	0.501 _{0.470-0.532}
LLaMA-3.1-8B-Instruct (SFT)	0.485 _{0.460-0.510}	0.278 _{0.249-0.307}	0.411 _{0.378-0.443}	0.479 _{0.448-0.509}
Qwen-2.5-7B (SFT)	0.486 _{0.461-0.510}	0.249 _{0.221-0.278}	0.363 _{0.335-0.394}	0.425 _{0.394-0.455}

set is stitched once, before training, by the *base* checkpoint of the model to be fine-tuned, so the fine-tuned model never generates its own training data. Only the training split is stitched; evaluation uses the enumerated reference reasoning. An example of a stitched reasoning trace can be found in the Supplementary Material.

Using supervised fine-tuning (SFT), we train two popular open-source base models, namely *Qwen-2.5-7B-Instruct* and *LLaMA-3.1-8B-Instruct*. Additionally, to test the marginal value of our reasoning dataset, we include *MedReason-8B*, a model based on *LLaMA-3.1-8B-Instruct* that was previously fine-tuned on a synthetic medical reasoning dataset. We updated each model with full-weight fine-tuning for three epochs on 8 NVIDIA H100 GPUs with a learning rate of 2e-5 and batch size of 256. All other hyperparameters are default according to the *verl* Python package implementation of SFT (*version v0.3.0.rc0*). We focus on 7–8B models to match existing open medical reasoning models (*MedReason-8B*, *m1-7b-23k*) and because models of this size can be run locally by health systems; larger models are left to future work.

3. Results

3.1. Dataset Quality: Physician Validation of Extracted Cases

We performed a human validation of the model-generated case prompts, diagnostic reasons, and final diagnoses on the MedCaseReasoning dataset. Our study included four U.S. board-certified

physicians reviewing a total of 100 cases. Each physician reported spending an average of 10-20 minutes per case. The figures below are the proportion of the 100 reviewed cases for which the reviewing physician answered “yes” to each question, with bootstrapped 95% confidence intervals; they are not an inter-rater agreement statistic, since each case was reviewed by a single physician. We found that 98% of the generated cases were free of hallucinations in the case prompt or diagnostic reasoning. We also report that 92% of final diagnoses were reported to have been faithful to the article and reasonably inferrable from the details of the case prompt. Finally, 93% of diagnostic reasoning steps were faithful to the case report and clinically relevant. This review assessed the faithfulness and relevance of the extracted reasoning, not its completeness: it did not ask whether every alternative discussed in the source article was captured. The results are found in the Supplementary Material.

3.2. Benchmark Results: Diagnostic Accuracy

We evaluate the following combination of frontier reasoning models and popular open-sourced models: *OpenAI o3*, *DeepSeek R1*, *QwQ-32B*, *MedReason-8B*, *m1-7b-23k*, *LLaMA-3.1-8B-Instruct*, and *Qwen-2.5-7B-Instruct*.

The models are each evaluated on pass@1, pass@5, and pass@10 accuracy, listed in Table 3 (NEJM CPC results in the Supplementary Material). Overall, we found that diagnostic reasoning remains a difficult task for top-performing LLMs. *OpenAI o3* performs significantly better than the rest of the models, with 64.5% pass@10 accuracy on the `MedCaseReasoning` test set, whereas *DeepSeek R1* achieves 48.0%. In comparison, *OpenAI o3* and *DeepSeek R1* achieve similar scores of 62.3% and 43.7% pass@10 accuracy on NEJM CPC, signaling the validity of `MedCaseReasoning` as an open-access alternative evaluation set for diagnostic cases.

We find that supervised fine-tuning significantly improves model performance across both test sets. For example, on the `MedCaseReasoning` test set, *MedReason-8B* improves by 31% (11.9 percentage points, from 38.2% to 50.1%) pass@10 accuracy, outperforming *DeepSeek R1*. Of note, this model also improves by 18% (5.3 percentage points, from 29.5% to 34.8%) on NEJM CPC and outperforms *QwQ-32B*. This provides evidence of the generalizability of `MedCaseReasoning` training data, as NEJM CPC consists of out-of-distribution, hand-crafted diagnostic cases that are often much longer in length.

3.3. Benchmark Results: Reasoning Recall

We evaluate six of seven models that provide reasoning traces on reasoning recall and find that the top model, *DeepSeek R1*, covers 64.2% of the reasoning steps provided in case reports. We also found that *MedReason-8B*, a model explicitly trained on synthetically generated medical reasoning traces, did not have substantially higher reasoning recall, and even had lower recall than *LLaMA-3.1-8B-Instruct*’s base model. However, after SFT on `MedCaseReasoning`, each of the base models improved significantly on reasoning recall. For example, *MedReason-8B* improved by 28% (11.5 percentage points) and *Qwen-2.5-7B-Instruct* improved by 50% (16.2 percentage points).

Qualitatively, the reference reasoning that base models missed was often candidate diagnoses never considered or exclusionary symptoms for common diagnoses (we leave quantifying these error types to future work). For example, one case report focused on a case of Adult-onset Still’s

disease (AOSD), which presented very similarly to Rocky Mountain Spotted Fever (RMSF). The model (*m1-7b-23k*) made a diagnosis for RMSF, missing reasoning about the patient presentation from the case report: “We feel that our patient’s symptoms were not due to RMSF, because the rash was salmon colored and worsened with the fever spikes.”. We found a significant correlation between model performance and reasoning recall (Pearson $r=0.710$, $p=0.0485$), indicating the value of measuring reasoning steps as a proxy for model performance itself. Furthermore, we also observe a significant correlation between the length of the model’s reasoning trace and the reasoning recall ($r=0.790$, $p=0.0196$). Since recall does not penalize extraneous content, and the scored trace is selected conditional on correctness, both correlations should be interpreted with caution (Section 5).

4. Discussion

MedCaseReasoning addresses a critical gap in current medical benchmarks: the assessment of diagnostic accuracy without scrutinizing the underlying reasoning process. SFT with the MedCaseReasoning training dataset demonstrates that smaller models, such as *Llama 3.1 8B* and *Qwen 2.5 7B*, can achieve diagnostic accuracy comparable to or exceeding that of larger models like *QwQ-32B* and *DeepSeek-R1* after SFT. While prior studies^(22,23) have explored learning reasoning from synthetic traces generated by more powerful models, our research is the first to demonstrate the efficacy of training directly on diagnostic reasoning derived from clinician-written case reports.

Compared to established benchmarks like MedQA, where leading models such as *GPT-4o* have already achieved over 90% accuracy, diagnostic performance on MedCaseReasoning currently peaks at 64.5%, and it correlates strongly with performance on NEJM CPC (Figure 2), while being open-access and substantially larger.

5. Limitations

Our study presents several limitations. First, in order to enable scalable, systematic evaluation, we use LLMs to score model adherence to reasoning traces and diagnostic accuracy. LLMs are known to hallucinate facts and exhibit certain biases, which may be propagated when LLMs are used as judges as well. Our validation of the reasoning judge relied on a single physician and 33 cases (89 reasoning statements); a larger sample annotated by multiple physicians, with inter-annotator agreement, would give a more precise estimate of judge reliability. Some case reports may lack sufficient detail for a definitive diagnosis or present trivial cases. The QA conversion process can also introduce variability, where case details may be inadvertently left out or hallucinated. While we implemented a clinician-validated filtering pipeline for the test set to ensure reasoning statements are grounded in the case presentation, some intractable cases (eg. ones where the diagnosis cannot be made without certain information) or trivial cases (eg. ones where the diagnosis is given away in the prompt) still persist. In particular, some reference statements depend on findings withheld from the case prompt (e.g., histopathology), which our filters do not explicitly remove, and extraction may be incomplete, since physicians checked faithfulness and relevance but not completeness. Newer extraction models may reduce both errors.

Second, MedCaseReasoning captures the case presentation at a single timestep before asking for a final diagnosis. It does not reflect the iterative, multi-stage nature of real-world clinical diagnosis, which involves refining differential diagnoses based on evolving information from tests,

imaging, and treatment responses. Case reports also over-represent rare or complex presentations, so performance should not be read as everyday diagnostic accuracy.

Third, our reasoning recall metric only captures clinical reasoning provided within case reports. Diagnostic reasoning is inherently subjective, and while our extensive training corpus aims to encompass diverse diagnostic standards, the alignment metric should be interpreted as adherence to observed clinical reasoning patterns across a diverse range of clinicians rather than an absolute, single gold standard. Reasoning recall also measures coverage only and does not penalize unsupported or extraneous reasoning; consistent with this, trace length correlates with recall ($r = 0.790$), and a precision-style or length-controlled metric would complement it. Scoring a correct generation whenever one exists also couples recall to pass@10; recall on a fixed generation would separate the two.

Fourth, some open-access case reports may appear in LLM pretraining data, which could inflate scores, although low frontier-model accuracy, enrichment for post-2020 reports, and agreement with the license-restricted NEJM CPC suggest memorization does not dominate.

The models we train and open source are for research purposes only and are not intended to be used for diagnosis of real patients. Accordingly, the misuse of these models risks inaccurate medical diagnoses and potential patient harm.

6. Conclusion

We introduced MedCaseReasoning, an open-access benchmark and training set of 13,989 diagnostic cases built from clinician-written case reports, with reference reasoning that lets models be evaluated on how they reach a diagnosis, not only on the diagnosis. Current reasoning LLMs leave substantial room for improvement on both: the best model reaches 64.5% pass@10 accuracy, and the best open model recalls 64.2% of the reference reasoning. Fine-tuning 7–8B models on reasoning traces derived from MedCaseReasoning improves both diagnostic accuracy and reasoning recall, and these gains transfer to the held-out NEJM CPC benchmark. Because the pipeline can be rerun on new case reports, MedCaseReasoning can grow over time.

Data and Code Availability

Dataset: <https://huggingface.co/datasets/zou-lab/MedCaseReasoning>

Code and supplement:

<https://github.com/kevinwu23/Stanford-MedCaseReasoning>

Institutional Review Board (IRB)

This research used only publicly available, open-access case reports and involved no patient recruitment or interaction; physician review used no identifiable data beyond the published literature.

References

1. AMA, AMA: Physician enthusiasm grows for health care AI undefined american medical association (2025).

2. E. Goh, R. Gallo, J. Hom, E. Strong, Y. Weng, H. Kerman, J. A. Cool, Z. Kanjee, A. S. Parsons, N. Ahuja, E. Horvitz, D. Yang, A. Milstein, A. P. J. Olson, A. Rodman and J. H. Chen, Large language model influence on diagnostic reasoning, *JAMA Network Open* **7**, p. e2440969 (2024).
3. D. McDuff, M. Schaeckermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni *et al.*, Towards accurate differential diagnosis with large language models, *Nature*, 1 (2025).
4. D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang and P. Szolovits, What disease does this patient have? a large-scale open domain question answering dataset from medical exams, *arXiv [cs.CL]* (2020).
5. D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song and J. Steinhardt, Measuring massive multitask language understanding, *arXiv* (2020).
6. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. A. Y. Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam and V. Natarajan, Large language models encode clinical knowledge, *Nature* **620**, 172 (2023).
7. Y. Zuo, S. Qu, Y. Li, Z. Chen, X. Zhu, E. Hua, K. Zhang, N. Ding and B. Zhou, MedXpertQA: Benchmarking expert-level medical reasoning and understanding, *arXiv [cs.AI]* (January 2025).
8. J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang and W. Yin, Large language models for mathematical reasoning: Progresses and challenges, *arXiv* (2024).
9. H. Ding, Z. Fan, I. Guehring, G. Gupta, W. Ha, J. Huan, L. Liu, B. Omidvar-Tehrani, S. Wang and H. Zhou, Reasoning and planning with large language models in code development, in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, (ACM, New York, NY, USA, August 2024).
10. V. L. Patel, J. F. Arocha and J. Zhang, Thinking and reasoning in medicine, *The Cambridge handbook of thinking and reasoning* **14**, 727 (2005).
11. Q. Jin, F. Chen, Y. Zhou, Z. Xu, J. M. Cheung, R. Chen, R. M. Summers, J. F. Rousseau, P. Ni, M. J. Landsman, S. L. Baxter, S. J. Al'Aref, Y. Li, A. Chen, J. A. Brejt, M. F. Chiang, Y. Peng and Z. Lu, Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine, *NPJ Digit. Med.* **7**, p. 190 (July 2024).
12. PubMed Central, PMC open access subset
<https://pmc.ncbi.nlm.nih.gov/tools/openftlist/>, Accessed: 2025-5-15.
13. Z. Kanjee, B. Crowe and A. Rodman, Accuracy of a generative artificial intelligence model in a complex diagnostic challenge, *Jama* **330**, 78 (2023).
14. Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen and X. Lu, PubMedQA: A dataset for biomedical research question answering, *arXiv [cs.CL]* (September 2019).
15. S. L. Fleming, A. Lozano, W. J. Haberkorn, J. A. Jindal, E. P. Reis, R. Thapa, L. Blankemeier, J. Z. Jenkins, E. Steinberg, A. Nayak, B. S. Patel, C.-C. Chiang, A. Callahan, Z. Huo, S. Gatidis, S. J. Adams, O. Fayanju, S. J. Shah, T. Savage, E. Goh, A. S. Chaudhari, N. Aghaeepour, C. Sharp, M. A. Pfeffer, P. Liang, J. H. Chen, K. E. Morse, E. P. Brunskill, J. A. Fries and N. H. Shah, MedAlign: A clinician-generated dataset for instruction following with electronic medical records, *arXiv [cs.CL]* (August 2023).
16. S. Bedi *et al.*, MedHELM: Holistic evaluation of large language models for medical tasks, *arXiv preprint arXiv:2505.23802* (2025).
17. B. K. Nayak, The significance of case reports in biomedical publication, *Indian J. Ophthalmol.* **58**, 363 (September 2010).
18. Gemini, Gemini: A family of highly capable multimodal models, *arXiv [cs.CL]* (December 2023).
19. E. Goh, R. J. Gallo, E. Strong, Y. Weng, H. Kerman, J. A. Freed, J. A. Cool, Z. Kanjee, K. P. Lane, A. S. Parsons, N. Ahuja, E. Horvitz, D. Yang, A. Milstein, A. P. J. Olson, J. Hom, J. H. Chen and A. Rodman, GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized

- controlled trial, *Nature Medicine* **31**, 1233 (2025).
20. E. Strong, A. DiGiammarino, Y. Weng, A. Kumar, P. Hosamani, J. Hom and J. H. Chen, Chatbot vs medical student performance on free-response clinical reasoning examinations, *JAMA Intern. Med.* **183**, 1028 (September 2023).
 21. T. Savage, A. Nayak, R. Gallo, E. Rangan and J. H. Chen, Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine, *npj Digital Medicine* **7**, p. 20 (2024).
 22. J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou and B. Wang, HuatuoGPT-o1, towards medical complex reasoning with LLMs, *arXiv [cs.CL]* (December 2024).
 23. J. Wu, W. Deng, X. Li, S. Liu, T. Mi, Y. Peng, Z. Xu, Y. Liu, H. Cho, C.-I. Choi, Y. Cao, H. Ren, X. Li, X. Li and Y. Zhou, MedReason: Eliciting factual medical reasoning steps in LLMs via knowledge graphs, *arXiv [cs.CL]* (April 2025).

Beyond ICD Codes: Fine-Tuning LLMs for In-Hospital Cardiac Arrest Identification from EHR Notes

Davy Weissenbacher[†], PhD

*Department of Computational Biomedicine, Cedars-Sinai Medical Center,
Los Angeles, CA, USA E-mail: davy.weissenbacher@cshs.org*

Jonathan Vo[†], MD

*Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System,
Los Angeles, CA, USA E-mail: Jonathan.Vo@cshs.org*

Audrey Uy-Evanado, MD

*Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System,
Los Angeles, CA, USA E-mail: Audrey.Evanado@csmc.edu*

Kyndaron Reinier, PhD

*Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System,
Los Angeles, CA, USA E-mail: Kyndaron.Reinier@csmc.edu*

Harpriya Chugh, BE, MSHS

*Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System,
Los Angeles, CA, USA E-mail: Harpriya.Chugh@cshs.org*

Kadiyala Vishnu, MD

*Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System,
Los Angeles, CA, USA E-mail: Vishnu.Kadiyala@cshs.org*

Mythreye Venkatesan, BE, MS

*Department of Computational Biomedicine, Cedars-Sinai Medical Center,
Los Angeles, CA, USA E-mail: Mythreye.Venkatesan@cshs.org*

Karen O'Connor, MS

*Perelman school of medicine, University of Pennsylvania,
Philadelphia, PA, USA
E-mail: karoc@pennmedicine.upenn.edu*

Sumeet S. Chugh[‡], MD

*Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System,
Los Angeles, CA, USA E-mail: Sumeet.Chugh@csmc.edu*

Graciela Gonzalez-Hernandez[‡], PhD

Department of Computational Biomedicine, Cedars-Sinai Medical Center,

[†]Both authors contributed equally to this work

[‡]These authors jointly directed this work.

Los Angeles, CA, USA E-mail: Graciela.GonzalezHernandez@csmc.edu

Background: Identification of in-hospital cardiac arrest (IHCA) through manual chart abstraction is the gold standard, but its time-consuming and resource-intensive, which naturally limits its practicality. Diagnostic codes provide an accessible automated alternative, but prior work has shown this method of extraction has poor sensitivity and poor positive predictive value. More accurate and scalable automated methods are needed to support IHCA surveillance and research.

Methods: We developed large language model (LLM)-based systems that incorporated physician-defined clinical knowledge through guideline-enriched prompting, supervised fine-tuning, and a correction procedure. The systems identified IHCA events and their locations from electronic health record notes. Performance was evaluated at the encounter level against physician-adjudicated chart review and compared with ICD-code-based identification.

Results: Compared with an ICD-code-based algorithm (precision 0.68, recall 0.91, F1 score 0.78), the supervised fine-tuned LLM offered substantially better performance for IHCA identification on a curated evaluation corpus, reaching a precision of 0.88, a recall of 1.00, and an F1 score of 0.94. When tested on a larger, unselected validation cohort of 45,525 encounters, a separate LLM configuration—relying on guideline-enriched prompting and a self-correction procedure rather than fine-tuning—achieved a precision of 0.79 for encounter-level IHCA identification, meaning most encounters flagged by the model were genuine IHCA cases. The fine-tuned LLM underperformed on this cohort most likely due to its inability to self-correct its decisions. Using a separate model for the correction step, we mitigated this issue and reached a high precision of 0.94.

Conclusion: Explicit integration of clinical knowledge through prompting and supervised fine-tuning enabled accurate identification of IHCA events and locations from routine clinical documentation. In a two-pass workflow, these systems could prioritize encounters for clinician adjudication, reduce manual screening requirements, and support more scalable IHCA surveillance and registry development.

Keywords: In-Hospital Cardiac Arrest, Sudden Cardiac Arrest, Information Extraction, Large Language Models, Generative Systems, Supervise Fine-Tuning

1. Introduction

In-hospital cardiac arrest (IHCA) affects approximately 300,000 patients annually in the United States, corresponding to 1.6 to 2.8 events per 1000 hospital admissions.¹ Survival to hospital discharge remains no higher than approximately 25%, underscoring the substantial mortality associated with these events.² IHCA is conventionally defined as cessation of cardiac activity in a hospitalized patient requiring chest compressions and/or defibrillation as part of cardiopulmonary resuscitation (CPR).³

Despite advances in resuscitation care, the incidence of IHCA has continued to increase, while improvements in survival have recently plateaued.^{1,4} These trends highlight the need

for accurate surveillance systems that can support benchmarking, outcomes research, and data-driven quality improvement. Accordingly, the American Heart Association’s “Ten Steps Toward Improving In-Hospital Cardiac Arrest Quality of Care and Outcomes” identifies the development of high-quality IHCA registries as a central component of efforts to improve resuscitation care.⁵

Current IHCA research relies heavily on prospective registries in which trained personnel manually abstract information from medical records and cardiac arrest flow sheets. The largest such registry in the United States is the American Heart Association’s *Get With The Guidelines–Resuscitation* registry.⁶ However, only 5–6% of US hospitals participate in the registry in a given year, limiting the representativeness and reach of registry-based surveillance.⁷ Expanding participation will require methods that preserve the clinical validity of expert adjudication while reducing the associated abstraction burden.

Outside prospective registries, IHCA cohorts are generally identified through manual chart review or through International Classification of Diseases (ICD) diagnosis and procedure codes extracted from the electronic health record (EHR).^{8,9} Manual review provides clinically detailed case ascertainment but requires substantial time and specialized expertise. ICD-based methods are considerably more scalable, but they have demonstrated limited sensitivity and positive predictive value for IHCA because of missing codes, miscoding, and codes carried forward from previous encounters.^{8,9} Regional and temporal variation in coding practices further limits their reliability for longitudinal surveillance and interinstitutional benchmarking.¹⁰

Clinical notes provide a richer representation of the events, interventions, and contextual information needed to determine whether an IHCA occurred. Natural language processing (NLP) and large language models (LLMs) have been applied to clinical text to identify cardiovascular phenotypes and risk factors, including heart failure, recurrent atrial fibrillation, and heart disease risk factors.^{11–14} To our knowledge, however, prior work has not addressed automated IHCA case identification from unfiltered clinical documentation. This task requires more than recognition of cardiac-arrest terminology: a system must apply clinically defined criteria, distinguish current events from historical or hypothetical mentions, determine whether resuscitative treatment occurred, and resolve whether the event began in an inpatient unit, the emergency department, or outside the hospital.

In this study, we investigate two complementary approaches for integrating physician-defined IHCA knowledge into LLMs. First, we explicitly encode clinical definitions and annotation rules through guideline-enriched prompting and few-shot examples. Second, we incorporate the same clinically adjudicated knowledge into the model parameters through supervised fine-tuning and evaluate and refine its predictions using external-model or self-correction procedures. Although these methods are widely used in this rapidly evolving field, our contribution lies in their systematic combination and evaluation for IHCA detection, particularly across curated and real-world cohorts.¹⁵ We develop and evaluate LLM-based systems for identifying IHCA events and their locations from unfiltered EHR notes, using physician-adjudicated chart abstraction as the reference standard and ICD-code-based identification as the principal comparator.

2. Materials and Methods

2.1. Data Compilation and Annotation

This study was reviewed and approved by the Cedars-Sinai Medical Center Institutional Review Board. We constructed two datasets for model development and evaluation. The first, termed the *Evaluation Corpus*, comprised 500 clinical encounters, defined as a complete episode of care from intake through discharge, from the Cedars-Sinai Medical Center electronic health record (EHR). An individual patient could contribute more than one encounter. For each encounter, the corpus included the available inpatient documentation, including emergency department (ED) notes, discharge summaries, and Code Blue notes recorded following cardiac arrest resuscitation. An encounter could contain multiple notes of the same type.

To obtain sufficient positive examples for model development, the Evaluation Corpus combined code-enriched and randomly sampled encounters. Of the 500 encounters, 47% contained at least one diagnosis or procedure code selected to identify a potential cardiac arrest, while the remaining 53% were randomly sampled encounters. Code presence was used only for cohort enrichment and was not treated as evidence that an IHCA had occurred. The codes used for sampling are listed in Table 1.

Table 1: Diagnosis and procedure codes used to enrich the Evaluation Corpus for potential cardiac arrest encounters

Code	Description
I49.01	Ventricular fibrillation
I49.02	Ventricular flutter
I46.2	Cardiac arrest due to underlying cardiac condition
I46.8	Cardiac arrest due to other underlying condition
I46.9	Cardiac arrest, cause unspecified
92950	Cardiopulmonary resuscitation
5A12012	Performance of cardiac output, single, manual
02QA0ZZ	Repair heart, open approach

Before annotation began, the two lead investigators iteratively developed annotation guidelines grounded in the Utstein definition of IHCA as pulselessness requiring chest compressions and/or defibrillation in a hospitalized patient.³ Two physicians (JV and AUE) independently applied the draft guidelines to a pilot sample of 50 notes. They adjudicated disagreements and refined the guidelines until no further modifications were required. The final annotation guidelines were previously published by Vo et al.¹⁶

Using these guidelines, JV and AUE annotated 3,822 notes from the 500 encounters. The physician annotations served as the reference standard. A note was labeled *IHCA* when the clinical text clearly documented a pulseless event treated with chest compressions and/or external defibrillation; otherwise, it was labeled *not IHCA*. Labels were based on whether resuscitative treatment was delivered rather than on whether treatment was clinically indicated or concordant with the patient’s code status. Thus, a pulseless event was labeled *IHCA* when

chest compressions and/or external defibrillation were performed, including when the patient had a pre-existing do-not-resuscitate or comfort-care order. A pulseless event for which resuscitation was not attempted was labeled *not IHCA*. A note was classified as positive for IHCA only when it documented both the occurrence of cardiac arrest or pulselessness and the delivery of resuscitative treatment, such as chest compressions or defibrillation. In the rare cases in which these two criteria were documented separately across multiple notes, the presence of multiple notes for the same patient generally ensured that at least one note documented both criteria together, thereby minimizing the risk of missing an IHCA.

Identifying all episodes of in-hospital pulselessness, including events without attempted resuscitation, represents an important but distinct surveillance objective relevant to end-of-life care and adherence to treatment limitations. Such events were outside the scope of this study, which was designed to align with Utstein-style IHCA registries focused on cardiac arrests for which resuscitation was attempted.³

For each note labeled positive for IHCA, we also annotated the location of the arrest. Location was categorized as: 1) in-hospital (IH), including critical care units, operating rooms, and cardiac catheterization laboratories; 2) emergency department (ED); and 3) out-of-hospital cardiac arrest for which treatment was initiated out of the hospital and continued on arrival to the emergency department (OH-ED). The third category, OH-ED (out-of-hospital cardiac arrest with continued treatment in the ED), is not conventionally included in standard IHCA definitions; its inclusion here was intended to facilitate identification of this subgroup and permit its exclusion in future research, as appropriate. For notes documenting multiple arrests, only the first arrest was labeled.

Following annotation of each individual note within an encounter, two physicians reviewed all notes for each encounter to adjudicate IHCA status and location at the encounter level. If there were multiple IHCA events in one encounter, the encounter-level location was adjudicated from the first event.

We randomly split the Evaluation Corpus into a training set (300 encounters, 60%), a development set (50 encounters, 10%), and a test set (150 encounters, 30%). The training set included 99 encounters with at least one IHCA event, comprising 172 notes labeled positive for IHCA. Of these 99 encounters, 6% were further classified as OH-ED (6 encounters, 12 notes), 17% as ED (17 encounters, 37 notes), and 77% as IH (76 encounters, 123 notes). The development set included 11 positive encounters, comprising 22 IHCA-positive notes. Of these encounters, 18% were classified as OH-ED (2 encounters, 5 notes), 9% as ED (1 encounter, 3 notes), and 73% as IH (8 encounters, 14 notes). Finally, the test set included 37 positive encounters, comprising 70 IHCA-positive notes: 8% of these encounters were classified as OH-ED (3 encounters, 6 notes), 16% as ED (6 encounters, 16 notes), and 76% as IH (28 encounters, 48 notes). Two physicians independently annotated the same 50 encounters for presence and location of IHCA, noting a Cohen's kappa of 0.951 for IHCA presence (almost perfect agreement) and 0.838 for IHCA location (strong agreement).¹⁷

Because the Evaluation Corpus was enriched with potential IHCA cases identified using diagnosis and procedure codes, its IHCA prevalence was higher than that expected in routine hospital data. We therefore conducted an additional evaluation using a year of encounters

from the ongoing Observational Study of Sudden Cardiac Arrest (OSCAR) cohort.¹⁸ (dated between January 1 and December 31, 2023 ($n = 45,525$)), including inpatient, emergency department, and observation encounters, as well as encounters containing a Code Blue note. Ambulatory encounters were excluded. No additional code-based enrichment or case preselection was applied.

The LLM was applied to all encounters in the 2023 OSCAR subset, and encounters predicted as IHCA were reviewed by two physicians (JV and AUE).¹⁸ For each model-flagged encounter, the physicians reviewed all flagged notes and determined whether the event met the Utstein definition of IHCA.^{3,18} Encounters meeting this definition were classified as true positives, whereas the remaining model-flagged encounters were classified as false positives. Because model-negative encounters were not systematically adjudicated, this analysis was designed to estimate precision rather than recall or overall classification accuracy.

2.2. LLM choice and prompt development

As an initial step, we developed a strong off-the-shelf, prompt-based system to establish a performance baseline before supervised fine-tuning. We selected Meta’s LLaMA 3.1-405B instruction-tuned model because it is an open-weight model available with strong general instruction-following capabilities that can be deployed within on-premises computing environment, although it requires substantial computational resources.¹⁹

We evaluated different approaches, beginning with two prompt-based baselines. In the first baseline approach (line 2 in Table 2), the LLM was given a role description, a concise statement of the task (classifying notes mentioning IHCA and location), and the note text. In the second baseline (line 3 in Table 2), we augmented this prompt with a summary of the clinical guidelines used for annotation, concise definitions of clinical concepts, and rule-based restatements of the guidelines to compensate for the model’s limited exposure to clinician-authored medical documents. We then enhanced this second baseline using few-shot and chain-of-thought prompting.²⁰ Few-shot prompting was implemented by adding a 10 training examples with correct IHCA and location labels, selected through iterative error analysis to highlight common false positives as well as representative true positives and true negatives. Chain-of-thought prompting, asking the LLM to briefly justify its predictions and cite supporting phrases, aka quotes, from the note, promotes step-by-step reasoning to improve explainability and encourage grounding.²⁰ Finally (line 4 in Table 2) added a self-correction step on top of this enhanced prompt. For each note, we first obtained the model’s classification and reasoning, then triggered a second review only for notes initially classified as IHCA, as most remaining errors involved false positives or misassigned locations. In this review, the model received the initial classification, location, reasoning, and note text, and was instructed to confirm or revise its decision using one of three location-specific prompts (IH, ED, or OH-ED).

2.3. Supervised full fine-tuning LLM

In the second step of our approach, we investigated training the model to further improve performance. We deliberately selected a small model so that full-parameter supervised fine-tuning could be conducted within our available computational resources. This allowed us

to study direct task adaptation by updating all model weights, while maintaining a training procedure that was computationally feasible and reproducible. We did not evaluate parameter-efficient fine-tuning methods such as LoRA or preference-optimization approaches such as RLHF, as these would have required additional method-specific hyperparameter tuning and experiments beyond the scope of this study. Such approaches may be more appropriate for larger models and remain important directions for future work.

With numerous open-source LLMs now available spanning different sizes and performance levels, we chose the Qwen 3 series²¹ for our experiments. This series includes both dense and mixture-of-experts models ranging from 0.6 to 235 billion parameters and is specifically optimized for reasoning. Computational and time constraints led us to select the smaller Qwen3-4B Instruct variant. We conducted fine-tuning using the Transformer Reinforcement Learning (TRL) library,²² chosen for its integration with Hugging Face Transformers.

To train the model, we presented each note and, using the same final prompt as for the system in line 3 of Table 2, asked the model to generate a JSON object indicating whether the note mentioned an in-hospital cardiac arrest and, if so, where it occurred (IH, ED, or OH-ED), along with supporting quotes from the text and a short explanation for its decision. We corrected the model’s output whenever it diverged from labels assigned by our physicians; for notes labeled as mentioning an in-hospital cardiac arrest, we provided a single standardized explanation—”The patient experienced a cardiac arrest and received treatment for it”—along with the supporting quotes as the correction. We fine-tuned the model for eight epochs on four NVIDIA A100 GPUs using a learning rate of 1×10^{-5} and an effective batch size of 16 examples per optimization step. The effective batch size was obtained using a per-device batch size of 2 and two gradient-accumulation steps. Training used BF16 mixed precision. We used TRL’s default token-level cross-entropy loss, leaving all other training parameters at their default values. We selected the final checkpoint produced at the end of training.

2.4. *Testing and validation*

We evaluated the performance of each baseline and the final system on the test set of the Evaluation Corpus at the encounter level. Encounter-level IHCA predictions were chosen over note-level predictions because clinical and research applications depend on whether a patient experienced IHCA during an encounter, rather than whether an individual note mentions IHCA, and because ICD codes also identify IHCA at the encounter level.

As a conventional NLP baseline, we compared the performance of our trained model against a simple baseline (line 5): the out-of-the-box Qwen3-4B Instruct model, prompted with our final prompt and our 10 few-shot examples. We also fine-tuned Bio_ClinicalBERT,²³ a BERT-based encoder pretrained on biomedical literature and MIMIC-III clinical notes, as a four-class note-level classifier. Because its 512-token context window was exceeded by a 20% of the notes, longer notes were divided into overlapping 512-token windows with a 128-token stride and window level predictions were aggregated to the note level. Note-level predictions were subsequently aggregated at the encounter level using the same procedure as for the other systems. The model was fine-tuned for eight epochs with an effective batch size of 16, a learning rate of 2×10^{-5} , BF16 precision, and inverse-square-root-frequency class weighting,

with the checkpoint selected by encounter-level macro F1 on the development set.

We compared models performance to performance of ICD codes commonly used to identify IHCA (ICD codes I49.01, I49.02, I46.2, I46.8, I46.9, CPT code 92950, and procedure codes 5A12012 and 02QA0ZZ). ICD performance was evaluated in two steps. To assess recall, these codes were applied to a gold-standard list of known IHCA encounters from the annotated corpus, and any positive encounters not captured by the codes were counted as false negatives. To assess precision, we applied the same codes to encounters from the larger OSCAR study between 7/1/2021 and 10/31/2021 (2021 OSCAR subset), and a physician adjudicated each candidate as IHCA or not; encounters adjudicated as not IHCA were counted as false positives.

We summarized binary encounter-level performance for IHCA detection using precision, recall, and F1 score: precision (equivalent to positive predictive value) denotes the proportion of predicted IHCAs that are true IHCAs, recall (equivalent to sensitivity) denotes the proportion of true IHCAs correctly identified, and F1 score denotes the harmonic mean of the two. Multiclass performance for IHCA location was assessed using a confusion matrix, with correct predictions appearing on the diagonal (see Figure 1).²⁴

3. Results

Table 2 details the performance of all systems. First, all generative systems achieve higher F1 scores than the ICD-based system (line 1) and Bio_ClinicalBERT (line 8) on the test set of the Evaluation corpus, regardless of the settings used to generate answers. The Bio_ClinicalBERT baseline (line 8) achieved an encounter-level precision of 0.62, recall of 1.00, and F1 score of 0.76 on the Evaluation corpus test set. In preliminary experiments, inverse-frequency class weighting yielded the same recall of 1.00 but substantially lower precision (0.29) and F1 score (0.45); inverse-square-root-frequency weighting was therefore used for the final Bio_ClinicalBERT baseline.

Second, prompt refinement produced a six-point gain in F1 for Llama 405B with few-shot prompting (line 3), compared with the baseline prompt (line 2), on the Evaluation corpus test set. Adding the self-correction step then reduced recall from 1.0 to 0.97 (line 3 vs. line 4), while precision increased from 0.86 to 0.88 on the Evaluation corpus and from 0.47 to 0.79 on the OSCAR corpus. Finally, full fine-tuning supervision produced the largest F1 gains: the best-performing system (line 6) achieved perfect recall on the Evaluation corpus test set and an F1 score of 0.94, compared to 0.92 for LLaMA-405B, a model 100 times larger. On the OSCAR corpus, precision for this system achieve a low 0.59 compared to 0.79 for LLaMA-405B. Note that, on the Evaluation corpus, although the fine-tuned Qwen3-4B model (line 6) outperformed LLaMA-405B (line 4), the difference did not reach statistical significance (paired bootstrap on binary note-level classification: 95% CI [-0.002, 0.097], $p = 0.068$). Despite a substantial overall sample size ($N=1203$), the relatively small number of positive cases ($n=70$) limited statistical power to detect differences in F1 score between systems. By contrast, on the OSCAR corpus, the LLaMA-405B (line 4) achieved significantly higher precision than the fine-tuned Qwen3-4B model (line 6): (paired bootstrap on binary note-level classification, restricted to notes flagged by at least one system: 95% CI [-0.142, -0.067], $p < 0.001$).

Table 2: IHCA detection performance at the Encounter-level: Evaluation and OSCAR 2023 Corpora. FS stands for few-shots, SFT for supervised fine-tuning, SC for self-correction, LC for Llama correction.

	Evaluation corpus			OSCAR 2023 corpus
	Prec.	Rec.	F1	Prec.
1. ICD Codes	0.68	0.91	0.78	—
2. Llama 405B - Basic Prompt	0.77	1.0	0.87	—
3. Llama 405B - Final Prompt + FS	0.86	1.0	0.93	0.47
4. Llama 405B - Final Prompt + FS + SC	0.88	0.97	0.92	0.79
5. Qwen 3 4B - Final Prompt + FS + SC	0.86	0.73	0.79	—
6. Qwen 3 4B - Final Prompt + SFT + SC	0.88	1.00	0.94	0.59
7. Qwen 3 4B - Final Prompt + SFT + LC	0.94	0.89	0.92	0.94
8. Bio-ClinicalBERT	0.62	1.0	0.76	—

		Predicted			
		None	OH-ED	ED	IH
Actual	None	111	0	0	2
	OH-ED	1	2	0	0
	ED	1	0	6	0
	IH	2	0	0	25

Evaluation corpus

		Predicted			
		None	OH-ED	ED	IH
Actual	None	45335	7	2	2
	OH-ED	1	11	0	0
	ED	4	8	40	3
	IH	8	0	5	99

OSCAR corpus

Fig. 1: Confusion matrices for encounter-level IHCA prediction by the fine-tuned Qwen3-4B model with Llama correction (line 7 in Table 2) on the test set of the Evaluation Corpus (left) and the OSCAR Corpus (right). Values along the diagonal indicate correct predictions; values in red indicate incorrect predictions. None refers to no IHCA. OH-ED refers to SCA that begins outside of the hospital with treatment continuing in the ED. ED refers to arrest that begins in the ED. IH refers to arrest that begins while admitted to the hospital.

4. Discussion

This study demonstrates that physician-defined clinical knowledge can be incorporated into LLM-based IHCA identification through both contextual and parametric adaptation. Guideline-enriched prompting, few-shot examples, and targeted self-correction substantially improved the performance of an off-the-shelf LLaMA 3.1-405B model, while supervised full fine-tuning enabled a much smaller Qwen3-4B model to achieve the best overall performance. On the Evaluation Corpus, the fine-tuned Qwen3-4B model achieved a precision of 0.88, recall of 1.00, and F1 score of 0.94. These findings extend our previous work on LLM-based IHCA identification by showing that task-specific supervised adaptation can encode clinically

adjudicated definitions and decision rules within a smaller model.¹⁶

The progression across prompt configurations also illustrates the value of explicitly representing clinical knowledge. The basic prompt required the model to classify a note using only a role description and task statement. Adding the physician-developed annotation guidelines, clinical definitions, selected examples, and evidence-grounded rationale instructions increased F1 from 0.87 to 0.93. The self-correction step produced a modest reduction in recall but improved precision, particularly in the lower-prevalence OSCAR cohort. These results suggest that explicit decision criteria and targeted review of likely error cases can improve discrimination between true IHCA events and clinically similar documentation, including historical arrest mentions, ED arrests, and out-of-hospital arrests with continued treatment in the ED.

Supervised fine-tuning produced an additional improvement over prompting alone. The fine-tuned Qwen3-4B model achieved an F1 score of 0.94, compared with 0.92 for the prompted LLaMA 3.1-405B model with self-correction. Although this difference did not reach statistical significance on the Evaluation Corpus, the finding is operationally relevant: a 4-billion-parameter model may be more feasible for secure local deployment than a 405-billion-parameter model. The result also indicates that task-specific clinical supervision can partially compensate for differences in model scale.

The OSCAR analysis addressed a different question from the curated Evaluation Corpus. Rather than estimating complete classification performance, it evaluated the precision of model-positive predictions in a year-long cohort without code-based enrichment. The prompted LLaMA model's precision increased from 0.47 to 0.79 after self-correction, indicating that the second-pass review substantially reduced false-positive screening results. The final precision of the fine-tuned Qwen3-4B model was lower with 0.59. Because model-negative OSCAR encounters were not systematically adjudicated, recall, overall accuracy, and the underlying IHCA prevalence cannot be estimated from this analysis. Accordingly, the OSCAR findings support the model's potential utility for prioritizing encounters for review but do not constitute external validation or a complete population-level performance assessment.

These findings have practical implications for IHCA surveillance. Manual abstraction remains necessary for high-validity registry construction but requires substantial clinical effort.² A two-stage workflow could use the model to screen routine documentation and present likely IHCA encounters, together with supporting text, for clinician adjudication. Such an approach could reduce the number of records requiring full review while preserving human oversight. However, the present study did not directly measure abstraction time, number needed to review, reviewer agreement with model-provided evidence, computational cost, or workflow integration. These operational outcomes should be evaluated prospectively before claims of reduced workload or improved registry participation can be confirmed. If validated across institutions, such systems could lower barriers to participation in IHCA surveillance and quality-improvement initiatives.⁵

The study also highlights limitations in the clinical knowledge available to general-purpose LLMs. Publicly available models are trained predominantly on nonclinical text, and their performance may be constrained by limited exposure to clinical documentation and institution-specific language.^{25,26} Guideline-enriched prompting partially addresses this limitation by sup-

plying relevant knowledge at inference time, while supervised fine-tuning incorporates task-specific knowledge into model parameters. Further adaptation using clinical documents and reinforcement learning with human-feedback may improve robustness across note types and documentation styles.^{27,28}

Several limitations warrant emphasis. First, all data were obtained from a single health system. Documentation conventions, Code Blue workflows, note templates, and terminology may differ across institutions and EHR platforms. Evaluation at independent health systems is therefore necessary before generalizability can be established.

Second, the Evaluation Corpus was intentionally enriched with encounters containing cardiac-arrest-related codes and included relatively few positive encounters in the test set. Although this corpus allowed us to estimate recall because all encounters were adjudicated, its sampling design and limited number of positive cases introduce uncertainty; therefore, the observed recall of 1.00 should be interpreted cautiously and may not generalize to the OSCAR population. Adjudicating a sufficiently large sample of model-negative encounters in OSCAR would be an important direction for future work.

Third, only model-positive encounters were adjudicated in the 2023 OSCAR analysis, limiting that evaluation to precision.

Fourth, the training targets used a standardized explanation rather than physician-authored reasoning tailored to each example. Consequently, the generated rationales and supporting quotations were not evaluated for completeness, faithfulness, or clinical usefulness and should not be interpreted as validated explanations. Richer supervision or human-feedback methods may improve evidence selection, but these approaches would require separate evaluation.

Fifth, the relatively poor performance of the fine-tuned Qwen3-4B with self-correction on the OSCAR corpus (line 4) was initially surprising. Our analysis revealed that the fine-tuned Qwen3-4B was largely unable to auto-correct its answers: it almost always confirmed its original decision, and in the rare cases where it did revise its output, the change was limited to the location of the cardiac arrest rather than the underlying classification. This resulted in a large number of false positives that persisted despite the self-correction step—errors that Llama-405B, by contrast, was able to correctly resolve during its own self-correction. We hypothesize that this reflects a loss of calibration following task-specific fine-tuning: as the model became increasingly specialized in detecting IHCA within the training distribution, it also grew overconfident in its predictions, which is consistent with catastrophic forgetting of the general self-critique behavior present in the base instruction-tuned model as supported by our series of experiments in.²⁸ To address this, we tested an alternative pipeline in which Llama-405B performs the correction step on the fine-tuned model’s outputs, rather than the fine-tuned model correcting itself. This pipeline improved overall performance on the OSCAR corpus, achieving the best precision of 0.94. On the Evaluation Corpus, it likewise improved precision to 0.94 without substantially reducing recall (0.89), maintaining comparable overall performance to the fine-tuned Qwen3-4B with self-correction, with an F1 score of 0.92.

Finally, the present results reflect retrospective model development. Prospective implementation would require secure data handling, access controls, audit trails, monitoring for

performance drift, and clearly defined clinician responsibility. The model should be used to support, rather than replace, physician adjudication.

5. Conclusion

Explicit integration of physician-defined clinical knowledge through guideline-enriched prompting, supervised fine-tuning, and targeted correction enabled accurate identification of IHCA events and their locations from EHR notes. A fully fine-tuned 4-billion-parameter model achieved the strongest performance to a substantially larger prompted model.

These findings support the use of locally deployable LLMs as screening tools for retrospective IHCA cohort construction. In a clinician-supervised workflow, the model could prioritize likely IHCA encounters for adjudication and reduce reliance on diagnosis and procedure codes alone. Validation across independent health systems, complete assessment in lower-prevalence cohorts, evaluation of model-provided evidence, and prospective measurement of abstraction efficiency are required before operational deployment or broad registry use.

Data Availability

Prompts are available upon reasonable request.

Code Availability

Code available upon reasonable request.

Competing Interests

The authors declare no competing interests.

Funding

This work was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM014731.

References

1. L. Wu, B. Narasimhan, K. Bhatia *et al.*, Temporal trends in characteristics and outcomes associated with in-hospital cardiac arrest: A 20-year analysis (1999-2018), *J Am Heart Assoc* **10**, p. e021572 (2021).
2. M. J. Holmberg, C. E. Ross, G. M. Fitzmaurice, P. S. Chan, J. Duval-Arnould, A. V. Grosses-treuer, T. Yankama, M. W. Donnino and L. W. Andersen, Annual incidence of adult and pediatric in-hospital cardiac arrest in the united states, *Circ Cardiovasc Qual Outcomes* **12**, p. e005580 (2019).
3. J. P. Nolan, R. A. Berg, L. W. Andersen, F. Bhanji, P. S. Chan, M. W. Donnino, S. H. Lim, M. H. Ma, V. M. Nadkarni, M. A. Starks, G. D. Perkins, P. T. Morley and J. Soar, Cardiac arrest and cardiopulmonary resuscitation outcome reports: Update of the utstein resuscitation registry template for in-hospital cardiac arrest: A consensus report from a task force of the international

- liaison committee on resuscitation (american heart association, european resuscitation council, australian and new zealand council on resuscitation, heart and stroke foundation of canada, interamerican heart foundation, resuscitation council of southern africa, resuscitation council of asia), *Resuscitation* **144**, 166 (2019).
4. J. Vo, F. L. Norby, P. Marano, Y. Matusov, K. Reinier, J. Ebinger, H. Halperin and S. S. Chugh, Management and prevention of in-hospital cardiac arrest: present and future, *npj Cardiovascular Health* **1**, p. 7 (2024).
 5. B. K. Nallamothu, R. Greif, T. Anderson, H. Atiq, T. B. Couto, J. Considine, A. R. De Caen, T. Djärv, A. Doll, M. J. Douma, D. P. Edelson, F. Xu, J. C. Finn, G. Firestone, S. Girotra, K. G. Lauridsen, C. K. Leong, S. H. Lim, P. T. Morley, L. J. Morrison, A. Moskowitz, A. Mulasari Sankardas, M. T. M. Mohamed, M. C. Myburgh, V. M. Nadkarni, R. W. Neumar, J. P. Nolan, J. Athieno Odakha, T. M. Olasveengen, J. Orosz, G. D. Perkins, J. K. Previdi, C. Vailancourt, W. H. Montgomery, C. Sasson and P. S. Chan, Ten steps toward improving in-hospital cardiac arrest quality of care and outcomes, *Circ Cardiovasc Qual Outcomes* **16**, p. e010491 (2023).
 6. M. A. Peberdy, W. Kaye, J. P. Ornato, G. L. Larkin, V. Nadkarni, M. E. Mancini, R. A. Berg, G. Nichol and T. Lane-Trullt, Cardiopulmonary resuscitation of adults in the hospital: A report of 14720 cardiac arrests from the national registry of cardiopulmonary resuscitation, *Resuscitation* **58**, 297 (2003).
 7. A. Uzendu, S. Girotra and P. S. Chan, What cardiac arrest registries can tell us about health differences and disparities, *Resusc Plus* **18**, p. 100614 (2024).
 8. C. DeZorzi, B. Boyle, A. Qazi, K. Luthra, R. Khera, P. S. Chan and S. Girotra, Administrative billing codes for identifying patients with cardiac arrest, *J Am Coll Cardiol* **73**, 1598 (2019).
 9. R. Khera, J. A. Spertus, M. A. Starks, Y. Tang, S. M. Bradley, S. Girotra and P. S. Chan, Administrative codes for capturing in-hospital cardiac arrest, *JAMA Cardiol* **2**, 1275 (2017).
 10. S. J. Nelson, Y. Yin, E. A. Trujillo Rivera, Y. Shao, P. Ma, M. S. Tuttle, J. Garvin and Q. Zeng-Treitler, Are icd codes reliable for observational studies? assessing coding consistency for data quality, *Digit Health* **10**, p. 20552076241297056 (2024).
 11. J. W. Cunningham, P. Singh, C. Reeder, E. S. Lau, S. Khurshid, X. Wang, P. T. Ellinor, S. A. Lubitz, P. Batra and J. E. Ho, Natural language processing for adjudication of heart failure in the electronic health record, *JACC Heart Fail* **11**, 852 (2023).
 12. E. H. Houssein, R. E. Mohamed and A. A. Ali, Heart disease risk factors detection from electronic health records using advanced nlp and deep learning techniques, *Scientific Reports* **13**, p. 7173 (2023).
 13. P. M. Marti-Castellote, C. Reeder, B. Claggett, P. Singh, E. S. Lau, S. Khurshid, P. Batra, S. A. Lubitz, M. Maddah, O. Vardeny, E. F. Lewis, M. Pfeffer, P. Jhund, A. S. Desai, J. J. V. McMurray, P. T. Ellinor, J. E. Ho, S. D. Solomon and J. W. Cunningham, Natural language processing to adjudicate heart failure hospitalizations in global clinical trials, *Circ Heart Fail* **18**, p. e012514 (2025).
 14. C. Zheng, M. S. Lee, N. Bansal, A. S. Go, C. Chen, T. N. Harrison, D. Fan, A. Allen, E. Garcia, B. Lidgard, D. Singer and J. An, Identification of recurrent atrial fibrillation using natural language processing applied to electronic health records, *Eur Heart J Qual Care Clin Outcomes* **10**, 77 (2024).
 15. X. Du, Z. Zhou, Y. Wang *et al.*, Performance and improvement strategies for adapting generative large language models for electronic health record applications: A systematic review, *Int J Med Inform* **205** (2025).
 16. J. Vo, D. Weissenbacher, K. Reinier *et al.*, Leveraging large language models to identify in-hospital cardiac arrest, *JACC: Clinical Electrophysiology* (2026).
 17. M. L. McHugh, Interrater reliability: the kappa statistic, *Biochem Med (Zagreb)* **22**, 276 (2012).

18. K. Reinier, H. S. Chugh, A. Uy-Evanado *et al.*, Observational study of sudden cardiac arrest risk (oscar): Rationale and design of an electronic health records cohort, *IJC Heart & Vasculature* **56**, p. 101614 (2025).
19. A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten and A. Vaughan, The llama 3 herd of models, *arXiv preprint arXiv:2407.21783* (2024).
20. J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le and D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, *Advances in neural information processing systems* **35**, 24824 (2022).
21. A. Yang, A. Li, B. Yang *et al.*, Qwen3 technical report (2025).
22. L. von Werra, Y. Belkada, L. Tunstall *et al.*, Trl: Transformer reinforcement learning <https://github.com/huggingface/trl>, (2020).
23. E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann and M. McDermott, Publicly available clinical bert embeddings, in *Proceedings of the 2nd clinical natural language processing workshop*, 2019.
24. M. Sokolova and G. Lapalme, A systematic analysis of performance measures for classification tasks, *Information Processing & Management* **45**, 427 (2009).
25. N. H. Shah, D. Entwistle and M. A. Pfeffer, Creation and adoption of large language models in medicine, *JAMA* **330**, 866 (09 2023).
26. S. Bedi, H. Cui, M. Fuentes *et al.*, Medhelm: Holistic evaluation of large language models for medical tasks (2025).
27. Y. S. Low, M. L. Jackson, R. J. Hyde *et al.*, Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems, *Digit Health.* **11** (2025).
28. D. Weissenbacher, M. Shabbir, I. M. Campbell, C. T. Berdahl and G. Gonzalez-Hernandez, Enhancing medical knowledge in large language models via supervised continued pretraining on clinical notes, *medRxiv* (2026).

Monopath DAGs: Structured Patient Trajectories from Clinical Case Reports

Anson Zhou^{1,*}, Aaron Fanous^{1,*}, Vasiliki Bikia^{1,2}, Yaowei Deng¹, Shlok Natarajan¹, Sonnet Xu⁶, Ank A. Agarwal¹, Akshat Garg¹, Noah Fanous³, Lyndon Huang³, Preston Tolbert⁵, Jonathan Luu⁴, Emily Alsentzer^{1,6,†}, Roxana Daneshjou^{1,7,†}

¹*Department of Biomedical Data Science, Stanford University, Stanford, CA, USA*

²*Institute for Human-Centered AI & Computer Science, Stanford University, Stanford, CA, USA*

³*Department of Medicine, University of Texas Dell Medical School, Austin, TX, USA*

⁴*Department of Psychiatry and Behavioral Sciences, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA*

⁵*Department of Otolaryngology–Head and Neck Surgery, LSU Health Shreveport, Shreveport, LA, USA*

⁶*Department of Computer Science, Stanford University, Stanford, CA, USA*

⁷*Department of Dermatology, Stanford University, Stanford, CA, USA*

**Equal contribution. †Co-senior authors.*

E-mail: aron7628@stanford.edu

Clinical machine learning has limited access to rare diseases, atypical treatment responses, and uncommon care pathways, many of which are documented in clinical case reports but remain difficult to model from free text. We introduce Monopath Directed Acyclic Graphs (DAGs), a structured representation that converts case reports into ordered patient states linked by explicit clinical transitions. We construct Monopath DAGs from 485 lung cancer case reports and evaluate their fidelity to the source narratives. Graph-based reconstruction achieves a ClinicalBERT BERTScore of $F_1 = 0.795 \pm 0.041$; human review finds 94% node semantic validity, 87% node temporal validity, 94% edge temporal validity, and 98% edge semantic validity. We next evaluate the representation in embedding and generation tasks. Node-text embeddings derived from Monopath DAGs achieve higher Calinski–Harabasz scores than embeddings of the original reports (136 vs. 78; difference 95% CI 34–79), although this analysis does not encode graph order or edges. In a blinded physician evaluation, graph-conditioned synthetic cases were rated approximately 0.45 points higher on a 5-point scale than controls generated from the same source report without the graph for timeline validity, clinical actionability, and perceived safe actionability ($p = 0.005$, $p = 0.005$, and $p = 0.001$, respectively). Clinical validity did not differ significantly between conditions. These results show that structuring case reports as explicit patient states and transitions preserves source-trajectory information and produces synthetic narratives that physicians judge more temporally coherent and actionable than those generated from a flat textual representation.

Keywords: clinical trajectories; directed acyclic graphs; synthetic clinical data; case reports; clinical language models; patient trajectory modeling

1. Introduction

Healthcare remains a challenging domain for AI model development. Although clinical AI models have increasingly achieved strong performance on retrospective benchmarks, these gains have failed to consistently translate to prospective clinical evaluation and routine deployment.^{1–3} Limited and inconsistent clinical data contribute to this translation gap, particularly for rare diseases and uncommon clinical presentations. As a result, synthetic data generation has become a promising technique for ML practitioners to expand coverage of these underrepresented cases.

Although synthetic data can expand the coverage of rare clinical cases, additional samples do not guarantee the preservation of informative qualities. A generated case may remain coherent and clinically plausible while omitting findings that alter diagnostic reasoning or subsequent actions. Clinical summarization studies have demonstrated this failure: compressed representations can preserve a high-level account of a patient’s medical history while excluding information that clinicians need for downstream decisions.^{4,5} Synthetic generation therefore depends on preserving information that constrains subsequent clinical decisions, not only representing information sufficient to maintain clinical validity.

Longitudinal EHRs introduce additional loss of representational fidelity across the clinical data pipeline. Clinical encounters encompass temporally-distributed observations, interventions, treatment responses, and outcomes, which clinicians selectively document, and downstream systems subsequently normalize, aggregate, or summarize for analysis. Each transformation can alter the temporal and contextual relationships between events, including treatment-response sequences, unresolved findings, and dependencies between prior observations and subsequent actions. Recent longitudinal EHR studies demonstrate that event-ordering itself carries information associated with later clinical outcomes.⁶ Consequently, a representation may retain sufficient information to characterize the current clinical state while failing to preserve the relationships that constrain subsequent clinical decisions.

Current evaluations of synthetic clinical data primarily assess distributional fidelity, task-specific utility, privacy risk, and broad measures of clinical plausibility.⁷ These criteria determine whether generated data reproduce statistical properties of observed cohorts, retain performance on predefined prediction tasks, or satisfy general clinical-validity constraints. They do not directly assess whether a synthetic case preserves the sequence of findings, interventions, and responses needed to discriminate between subsequent clinical actions. As a result, two cases may achieve comparable validity or utility scores while materially differing in temporal coherence, actionability, and the safety of the decisions they support. Clinical representations can therefore be measured by clinicians’ perceived ability to act on them.

Narrative case reports provide a useful setting in which to study this gap. Clinicians use case reports to document unusual presentations, diagnostic workups, interventions, treatment responses, and longitudinal disease progression. These reports often retain the sequence of findings, decisions, and responses that condensed EHR representations omit, while disproportionately representing rare or diagnostically challenging cases. However, case reports encode these relationships in free text. Temporal dependencies, decision points, and transitions between clinically-meaningful states often remain implicit within the narrative, limiting

systematic extraction and large-scale modeling.

Although case reports may describe competing diagnoses, concurrent findings, and alternative management considerations, they ultimately document a single observed clinical course. Therefore, we can represent each case as a directed path through clinically-meaningful patient states and transitions. We retain this path within a directed acyclic graph (DAG) structure so that individual states and transitions remain explicitly addressable and the representation can support graph operations such as sub-trajectory selection. This study uses the complete observed trajectory rather than selected subgraphs, and tests whether conditioning generation on it changes how clinicians judge the resulting cases.

We present MONOPATH DAGs as structured intermediates for synthetic case generation. The framework extracts clinically-relevant states and transitions from case-report narratives, organizes them into temporally-ordered trajectories, and conditions generation on the resulting graph structure. This design allows us to compare synthetic cases generated with and without explicit trajectory structure while holding the underlying source material constant.

We evaluated this framework on 485 curated lung cancer case reports and 200 rare-disease cases across four disease groups. For each report, we generated two synthetic case summaries: one conditioned on the report’s graph and one generated from the report without the graph (the control). Physicians rated each summary on clinical validity, timeline validity, clinical actionability, clinically safe action, and language appropriateness. This evaluation directly tests whether cases with similar clinical validity can still differ in the information they preserve for subsequent clinical decision-making.

This work makes four key contributions:

- **A decision-relevant evaluation framework** for synthetic clinical data that distinguishes clinical validity from timeline validity and clinical actionability and clinically safe action.
- **Evidence that clinical validity ratings can understate differences in clinical actionability:** graph conditioning raised clinical actionability by about 0.45 points on a 5-point scale, whereas the difference in clinical validity (+0.21, 95% CI −0.05 to 0.48) was smaller and not significant in the primary model.
- **A structured trajectory representation** that extracts ordered, ontology-grounded Monopath DAGs from free-text case reports, representing each case as a sequence of patient states and transitions.
- **A synthetic data generation pipeline and evaluation dataset** spanning 485 lung cancer case reports and 200 rare disease cases across four disease groups.

2. Related Work

Clinical Narrative Information Extraction. Information extraction from clinical text has progressed from identifying entities to recovering the relations among them. Rule- and vocabulary-based systems such as cTAKES and MetaMap normalize entities to controlled terminologies. Statistical and transformer-based methods extend this to temporal events and their ordering, including semi-supervised CRFs for temporal event detection in EMRs,⁸ ClinicalBERT⁹ and BioGPT,¹⁰ and GraphTrex on i2b2 TempEval.¹¹ Other methods extract

cause–effect relations through dependency patterns.¹² Other work targets case reports directly, through LLM-based extraction,¹³ graph-based representations of cardiovascular case reports in CREATE,¹⁴ and the 167,000 case summaries of PMC-Patients for retrieval.¹⁵ These approaches recover entities, relations, temporal events, or condensed summaries, but do not represent a case as the sequence of patient states it describes and the transitions between them. In contrast, we represent each report as an ordered sequence of patient states, with the transition between each pair of consecutive states stated explicitly.

Patient Trajectory Modeling. Trajectory models represent a patient’s course as a sequence of structured events. Sequence models such as RETAIN,¹⁶ BEHRT,¹⁷ and Med-BERT¹⁸ treat diagnosis codes or visits as tokens in an ordered sequence, and graph-based methods such as GraphCare¹⁹ add knowledge-graph structure over codes and visits for prediction. In these models a trajectory is a sequence of coded events drawn from structured records. They need coded inputs that case reports lack, and they omit the states and decisions a narrative describes. Few studies recover such trajectories from case reports.

Synthetic Patient Data Generation. Researchers generate synthetic patient data to offset data scarcity and privacy limits. Rule-based simulators such as Synthea²⁰ encode disease progression by hand, while learned generators synthesize longitudinal records with GANs and encoder–decoder models (EHR-M-GAN,²¹ EHR-Safe²²), hierarchical Transformers over code sequences (HALO²³), or language models for discharge summaries.²⁴ These generators are typically evaluated on distributional fidelity, downstream utility, and privacy,⁷ criteria that do not test whether a generated case preserves the information needed for the next clinical decision. Generators learned from cohorts also see few examples of the rare and atypical trajectories that case reports document.

Here, we recover state-level trajectories from case narratives, condition generation on them, and ask physicians whether the generated cases support the next clinical decision.

3. Methods

3.1. Modular Framework for Patient Trajectory Graph Extraction

We present a modular framework for converting free-text clinical case reports into structured patient trajectories represented as *Monopath Directed Acyclic Graphs (Monopath DAGs)*.

3.2. Task Definition

We represent each clinical case report $d_i \in \mathcal{D}$ as a directed acyclic graph $G_i = (V_i, E_i)$. Each node corresponds to a patient state, defined as a snapshot of the patient as perceived by the clinician at a given point in the trajectory. States reflect the clinical information available at that point rather than an assumed complete representation of the patient’s underlying condition. Nodes are annotated with free-text descriptions, structured clinical attributes, and, when available, explicit timestamps. Directed edges represent transitions between states and are labeled by transition type and the clinical entities associated with each transition. See

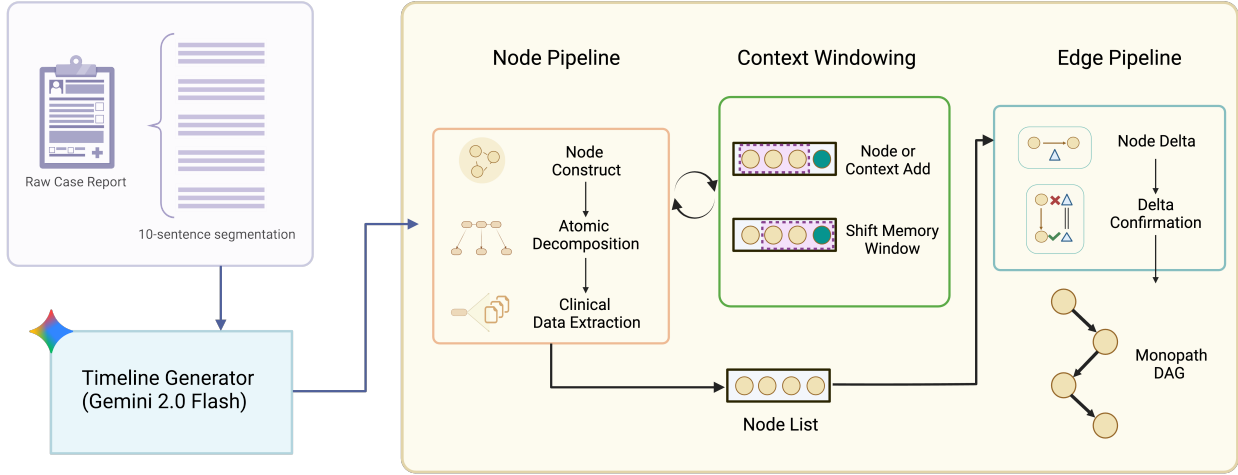


Fig. 1: DAG generation pipeline for patient trajectory modeling.

Supplement S1 for the complete formal definition of patient states and graph structure.^a

3.3. Patient Trajectory Generation

Trajectory construction follows three structural constraints: (i) the graph is acyclic, reflecting the forward progression of clinical time; (ii) each node has at most one incoming and one outgoing edge, and the graph is connected; and (iii) nodes are temporally ordered using explicit timestamps when available and inferred narrative order otherwise. These constraints define the Monopath representation and reduce ambiguity during trajectory construction. Topology validation flags graphs that violate these constraints (Section 5).

Preprocessing & Timeline Extraction. Each case report is retrieved from PubMed Central as UTF-8 HTML, with non-narrative boilerplate removed. We segment the remaining narrative into 10-sentence blocks and pass them to the timeline generator $T_i = \text{PatientTimeline}(p_1, \dots, p_k)$. In pilot runs, we judged qualitatively that Gemini 2.0 Flash extracted timelines more consistently than Llama 3.1 8B and 70B at acceptable cost. All language-model steps use Gemini 2.0 Flash at $T = 0.2$ with an approximately 128k-token context window.

Node Construction. We divide each patient timeline T_i into four-sentence blocks c_j and process them sequentially to construct patient states. For each block, we generate a node

$$v_j = \{\text{node_id}, \text{step_index}, s_j, a_j, t_j\},$$

where s_j is the free-text state description, a_j contains structured clinical attributes, and t_j is an optional timestamp. To preserve local temporal context, each node v_j is generated using a compressed representation of the three preceding states.

^aSupplementary material, prompts, code, and data:

<https://github.com/DaneshjouLab/MonopathDAGs/tree/main/supplement>

We recursively decompose each state description s_j into atomic clinical facts, defined as statements containing a single clinical assertion about the patient state. We then map these facts to structured ontology terms. Facts that cannot be mapped are retained as free text to preserve information. We merge nodes with matching identifiers (appending new content and combining clinical data), apply a final model pass to remove duplicate states, and reindex the sequence (Supplement S3). We do not model negation or uncertainty explicitly; temporality is carried by node order, with timestamps only when stated.

Edge Construction. Given the final node sequence $\{v_1, \dots, v_n\}$, we construct an edge between each adjacent pair (v_i, v_{i+1}) by identifying the clinical change between states. Each edge is represented as

$$e_{i \rightarrow i+1} = \{\text{edge_id}, \tau_{ij}, \phi_{ij}\},$$

where τ_{ij} is a free-text transition description and ϕ_{ij} is a structured event containing the trigger type, clinical domain, and associated ontology terms. We use three few-shot demonstrations (Supplement S5) to keep transition labels consistent.

Graph Serialization & Metadata Tracking. Each patient trajectory is serialized as a JSON directed graph following a consistent schema with ontology-grounded labels and structured metadata (Supplement S3). We enforce schema compliance during construction while deferring full structural validation for downstream evaluation.

3.4. Lung Cancer Case Report Dataset

We curated a dataset of full-text lung cancer case reports from PubMed Central (PMC). Using the Entrez API, we searched for open-access case reports published from 2019 onward using lung cancer-related terms, including lung cancer, lung carcinoma, non-small cell lung cancer (NSCLC), small cell lung cancer (SCLC), mesothelioma, pulmonary carcinoma, and bronchogenic carcinoma, matched against title and abstract fields. This search yielded 550 open-access case reports.

4. Evaluation and Utility

4.1. Graph Construction and Fidelity

We assess graph fidelity with human review, BERTScore, and structural checks.

Human Graph Review. For each edge and node of a random sampling of 30 graphs, two graduate students compare each generated graph with the source case report. Reviewers assess whether each node and edge preserves the patient states represented in the source case, the clinically relevant information contained within those states, and the temporal ordering of the trajectory (Supplement S4). The graduate students made 568 annotations of 445 elements (243 nodes, 202 edges) from 30 graphs; on the 123 elements annotated twice, agreement on G1 (correct order) and G2 (accurate content) was 94% and 92%. Cohen’s κ is 0.73 for G1; for G2, nearly all ratings were positive, which makes κ uninformative.

LLM-Based Narrative Reconstruction and BERTScore Evaluation. We reconstruct the input case narrative from the generated graph using Gemini 2.0 Flash. The model receives a graph’s nodes and edges and is prompted to generate a coherent textual summary that

aims to mirror the structure and content of the original case. The reconstructed narrative $\hat{R}$ is compared to the original case report R using BERTScore,²⁵ which measures semantic similarity via contextualized embeddings from ClinicalBERT. Reconstruction-based BERTScore captures presumed semantic coverage between the graph and the source narrative through the distance, it does not explicitly detect hallucinated content or subtle clinical inaccuracies, motivating complementary human evaluation. We calculate BERTScore with Bio_ClinicalBERT⁹ (layer 12, without idf or rescaling), calculated on the full texts: each text is embedded in consecutive 510-token windows, as in Section 4.2, and all tokens enter the greedy matching. Scores may partly reflect self-consistency, since one model constructs and reconstructs the graphs, or the model’s prior exposure to published reports.

Topology Validation. We sanity checked the structural integrity of the patient graphs through topological audits. We checked acyclicity by applying depth-first search, count connected components, and flag malformed graphs.

4.2. *Clustering of Patient Trajectories*

We embedded graphs and case reports with ClinicalBERT. A graph was represented by the mean of its node-text embeddings (node order and edges not encoded); a report, by the mean embedding of consecutive 510-token windows, so that no text was truncated (case presentations average 1,137 ClinicalBERT tokens, median 967, and 73% exceed 512). To test whether edges add information, we also embedded each edge description with the same model and represented each graph by the mean over all node and edge embeddings. We analyzed embeddings in 768 dimensions and after PCA or UMAP reduction to 10, and assessed cluster structure with K-means ($k \in [2, 10]$, selected by silhouette²⁶) and the Calinski–Harabasz (CH) index.²⁷ For the 484 patients with metastasis status, annotated manually for each report, we computed CH with patients partitioned by status, and the cross-validated AUC of logistic regression predicting status from each embedding. 95% CIs are from patient-level bootstrap resamples; settings are in Supplement S9.

4.3. *Synthetic Trajectories and Clinical Expert Review*

Synthetic Generation. We prompted a language model to generate 77 narrative vignette pairs from conditioning on the source or from the graph. The control extracts a list of clinical events from the full report in one pass and writes a narrative from it; the graph-conditioned narrative is generated from the ordered nodes and edges of G_i . Both use nearly identical prompting, and the same case per narrative pair (Supplement S5).

Clinician-centered evaluation criteria

- Q1. (Clinical validity):** Baseline check that case actions are medically sound.
- Q2. (Timeline validity):** Assesses whether the case progression follows a clinically appropriate order, capturing temporal reasoning.
- Q3. (Clinical actionability):** Measures whether the physician judges the case specific and granular enough to guide the next clinical decision.
- Q4. (Clinically safe action):** Measures whether the physician judges the information suffi-

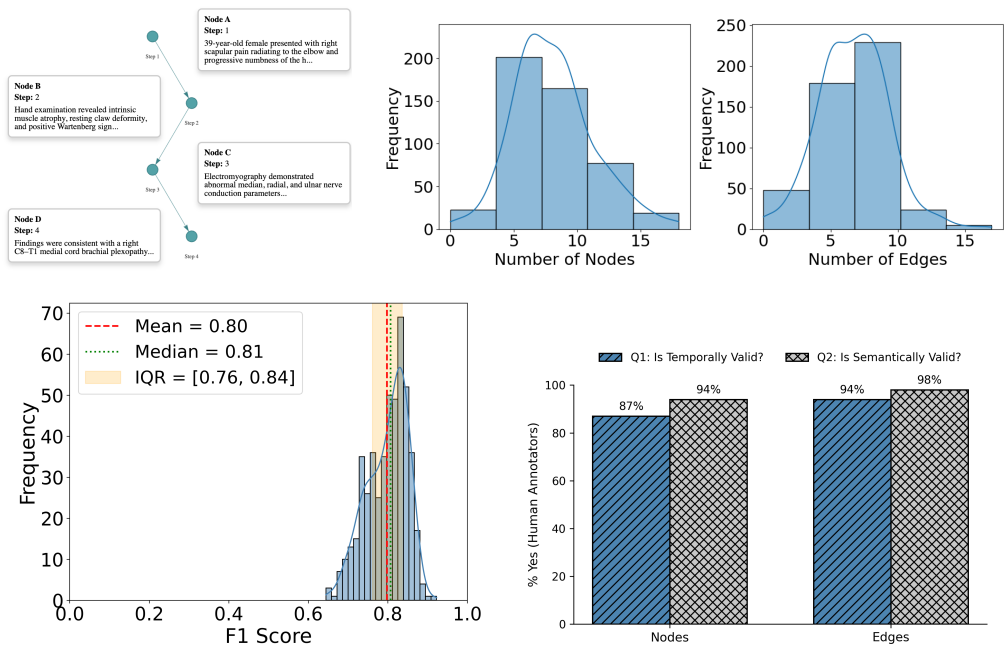


Fig. 2: **Monopath DAG trajectory representations and evaluation.** (a) Schematic of a Monopath DAG representing a patient case. (b) Distribution of node and edge counts across constructed graphs. (c) BERTScore F1 similarity between reconstructed narratives and original texts. (d) Human evaluation comparing graph fidelity and ordering accuracy.

cient to make a safe decision with confidence, reflecting the physician’s decision certainty.

Q5. (Appropriate language): Evaluates whether the narrative uses medical language appropriately, avoiding generic or ambiguous phrasing.

Clinical Expert Review. Resident physicians (PGY-1 to PGY-3), blinded to narrative generation conditions, rated each narrative on 5-point Likert scales. They receive cases in the same order, to ensure overlap coverage but generation conditioning order is randomized. Three of four physicians contributed 227 ratings (105 graph-conditioned, 122 control) across 76 cases: in 98 physician–case combinations they rated both versions (97 with complete answers), and in 31 they rated one version only (24 control, 7 graph-conditioned). The fourth rated one narrative and is excluded. We fit a linear mixed-effects model to each question, with crossed random intercepts for physician and case to absorb harsh raters and hard cases, and applied Holm²⁸ correction across the five questions. The graph-conditioned arm passes through more processing steps than the control, so the two arms do not receive equal input. The five questions (Q1–Q5), reviewed by medical experts, cover clinical validity, timeline validity, clinical actionability and clinically safe action, and language. Q1 assesses medical correctness in isolation, whereas Q4 evaluates whether the information provided supports confident and safe decision-making under clinical uncertainty (Supplement S4).

5. Results and Discussion

From the initial set of 550 case reports, we remove non-English entries and case series, resulting in a curated dataset of 485 individual narratives. Semantic fidelity of the graph-derived reconstructions is evaluated using BERTScore, yielding a mean F1 of 0.795 (95% CI 0.791–0.798), median 0.802, interquartile range [0.770, 0.824], range 0.643–0.879, and standard deviation 0.041. Reconstructions remain semantically close to their sources, although BERTScore does not detect omissions or clinical errors. Precision and recall components of BERTScore are 0.807 and 0.786, respectively (Supplement S2). On average, graphs contained 8 ± 3 nodes and 7 ± 3 edges. Node temporal validity is 87% and node semantic validity 94%; edge temporal validity is 94% and edge semantic validity 98%, indicating that most reviewed elements (30 graphs) were correctly ordered and accurate to the source (Figure 2d). Most reviewed nodes and edges were therefore accurate to the source and correctly ordered; node ordering was the least accurate of the four measures.

Graph embeddings yielded higher Calinski–Harabasz indices than case-report embeddings in every space: 136 (95% CI 119–156) vs. 78 (69–90) in the original space, 224 (198–258) vs. 147 (127–171) after PCA, and 738 (649–841) vs. 358 (318–408) after UMAP, with the difference excluding zero in each case. Both representations separated metastatic from non-metastatic patients to a similar degree (CH 10.6 vs. 7.0; difference 3.6, 95% CI -1.6 to 10.7), and the AUC for predicting metastasis status was slightly lower for graphs, not significantly different (0.73 vs. 0.77; difference -0.04 , 95% CI -0.09 to 0.01). Graphs reduce each report (mean 1,137 tokens) to about eight states. When we embedded each edge description with the same model and added the edges to the nodes, AUC for metastasis status rose by 0.01–0.03 in every cross-validation split and setting we tried, but the patient-level 95% CI for the difference included zero (-0.01 to 0.06). With edges, AUC approached that of the full text (0.755 vs. 0.769), and edges alone scored slightly below nodes. Adding edges lowered cluster compactness (CH 101 vs. 136) (Supplement S9). Graph embeddings therefore form more compact clusters than the report text, while metastasis status is predicted with similar accuracy from both, although each graph contains about eight states. Adding edges raised the AUC slightly but lowered compactness.

In a blinded evaluation, we analyzed the three of the four participating resident physicians (PGY-1 to PGY-3) who rated both versions of cases, comparing control-generated and graph-conditioned synthetic cases across the five clinical criteria (Figure 3, Table 1). PGY-level physicians were selected because they routinely perform frontline clinical reasoning and decision-making, making them well positioned to assess timeline coherence, actionability, and safety in narrative cases. We did not detect a difference in clinical validity (Q1; mixed-model estimate $+0.21$, 95% CI -0.05 to 0.48, $p = 0.12$), whereas graph-conditioned cases were rated higher on timeline validity (Q2, $+0.46$, 0.18 to 0.73, $p = 0.005$), clinical actionability (Q3, $+0.45$, 0.17 to 0.72, $p = 0.005$), and clinically safe action (Q4, $+0.45$, 0.21 to 0.68, $p = 0.001$; means 2.74 vs. 2.31, both below the scale midpoint, with 19% vs. 13% of ratings 4–5). Excluding physician A or B removes the significance of timeline validity and both actionability measures after Holm correction; excluding physician C does not, and without C, clinical validity also differs ($+0.36$, $p = 0.008$; Supplement S8). Agreement between physicians on individual ratings

was low (Spearman ρ -0.10 to 0.37). Language appropriateness (Q5) did not differ significantly after correction ($p = 0.058$). Graph conditioning therefore raised ratings of timeline validity, clinical actionability, and clinically safe action by about 0.45 points on a 5-point scale, while the difference in clinical validity was smaller and not significant. Ratings of clinically safe action stayed below the scale midpoint in both conditions, the effects depended on physicians A and B, and agreement between physicians was low. Although the binary evaluation schema (G1: order, G2: accuracy) provides only a coarse-grained view of graph fidelity, it is designed to complement automated semantic metrics such as BERTScore with a minimal human check that could be performed within the constraints of clinician time. This design emphasized inter-reviewer agreement and feasibility, but it does not capture node informativeness or the clinical validity of edges on its own. We consider this a key limitation of the current evaluation and plan to extend our framework with more fine-grained protocols, including explicit node-level assessments and validation of transition plausibility. As noted in Section 4.3, clinician evaluation compared synthetic narratives with control narratives on textual criteria; physicians rated narratives only; they did not view the graphs or compare formats. In this study, graphs are used as structured intermediates to support generation, rather than as interfaces intended for direct clinical decision-making.

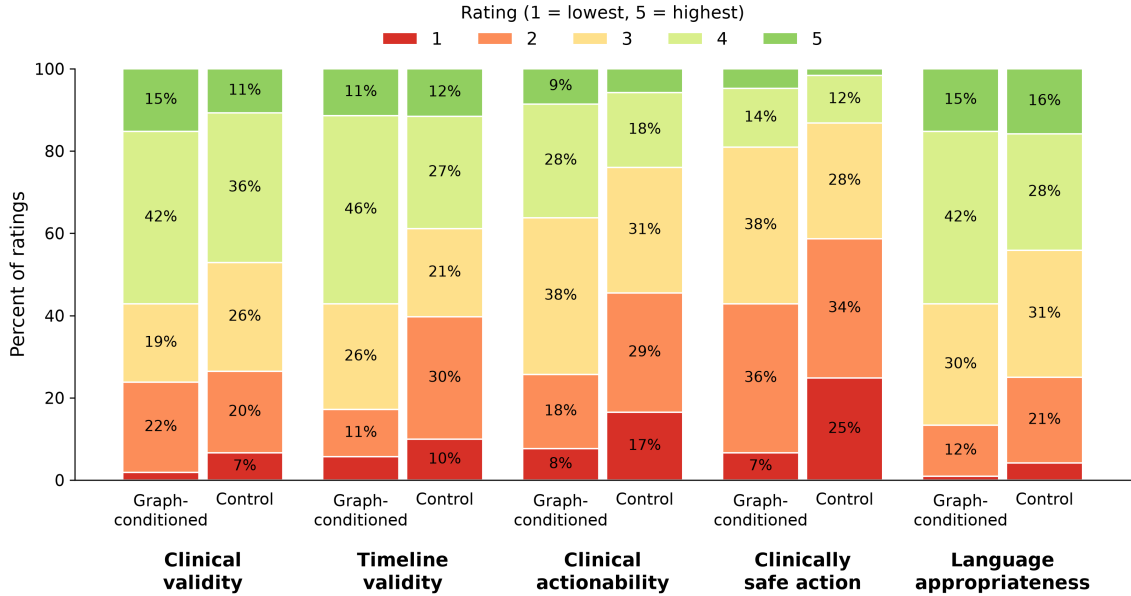


Fig. 3: Physician ratings of graph-conditioned and control narratives.

6. Exploratory Application to Rare Diseases

While our primary experiments focus on lung cancer—chosen for its prevalence in case report literature, detailed longitudinal trajectories, and heterogeneous treatment paths—the pipeline is agnostic to disease.

Table 1: Linear mixed-effects estimates (graph-conditioned – control) with crossed random intercepts for physician and case, across five Likert-scale questions. Four physicians took part; the 227 ratings analyzed, across 76 cases, came from the three physicians who contributed paired ratings. CIs are unadjusted; p values are Holm-adjusted across questions ($*p < 0.05$).

Question	Estimate	95% CI	p
Clinical validity	+0.21	[−0.05, 0.48]	0.117
Timeline validity	+0.46	[0.18, 0.73]	0.005*
Clinical actionability	+0.45	[0.17, 0.72]	0.005*
Clinically safe action	+0.45	[0.21, 0.68]	0.001*
Appropriate clinical language	+0.28	[0.03, 0.53]	0.058

6.1. Rare Diseases.

We conducted additional experiments on four rare diseases across body systems (Gaucher, Wilson’s, Charcot–Marie–Tooth, and Retinitis Pigmentosa), using 50 case reports per disease (200 total, see Supplement S2). These conditions were selected because they span distinct organ systems, disease progression patterns, and diagnostic pathways. We report this as an exploratory analysis of clustering only (Supplement S2), without human or downstream evaluation; it does not establish generalization or clinical utility.

7. Conclusion and Future Work

Monopath DAGs provide a structured intermediate for converting case reports into synthetic clinical narratives. The graphs are faithful to the source on most reviewed elements, yield more compact embeddings than the original text, and support synthetic narratives that physicians align more with timeline validity and clinical actionability and perceived clinically safe action. Clinical validity did not differ significantly. These findings suggest that evaluating synthetic cases on clinical plausibility alone can miss differences in meaningful components that lead to clinical decision making.

Limitations. Our evaluation is limited to recent, English-language lung cancer case reports and may not generalize across specialties, languages, or narrative conventions. Monopath DAGs enforce a single dominant clinical thread, collapsing concurrent events, branching trajectories, and cyclic complications. The extraction pipeline also depends on general-purpose LLMs and retains errors in negation, temporal ordering, and dosage extraction that are not fully captured by BERTScore or our limited human evaluation. Synthetic trajectories were evaluated through small-scale expert review, without model-leakage audits or downstream task benchmarking. We also do not explore privacy leakage from the case reports themselves. The single-path representation trades expressivity for a simple temporal structure that can be checked against the source. We did not quantify the information lost to the single-path assumption. The generation comparison does not separate segmentation, ordering, and composition, and the two arms differ in processing steps and prompt length. Case reports are curated, single-author narratives of unusual courses, whereas EHR trajectories are fragmented across

many notes and authors; we have had some success testing this on synthea data, but not at a larger scale and note that case reports are quite different from EHR knowledge.

Future Work. We plan to extend Monopath DAGs to *Multipath DAGs* with controlled branching to represent clinical uncertainty, reversible events, and alternative pathways. We will also investigate biomedical language models and graph-imputation methods to improve extraction of under-specified clinical states. These extensions could support trajectory-informed outcome prediction, counterfactual synthetic generation, cross-disease benchmarking, and trajectory-aware case matching for clinical decision support.

References

1. Why high scores do not mean application readiness for health ai, *Nature Medicine* (2026).
2. R. M. W. W. Tseng, L. C. Ong, J. H. L. Goh, Y. Chen, T. Chen, E. Lum and Y.-C. Tham, Prospective real-world implementation of deep learning systems in healthcare: a systematic review guided by implementation science, *npj Digital Medicine* **9**, p. 172 (2026).
3. Show us the evidence for the value of medical ai, *Nature Medicine* **32**, 1163 (2026).
4. J. D. Oliveira, H. D. P. Santos, A. H. D. P. S. Ulbrich, J. C. Couto, M. Arocha, J. Santos, M. M. Costa, D. Faccio, F. O. Tabalipa and R. F. Nogueira, Development and evaluation of a clinical note summarization system using large language models, *Communications Medicine* **5**, p. 376 (2025).
5. A. X. You, N. M. Kahl, A. Patel, J. Cao, E. M. Castillo, P. S. Bedmutha, T. Chan, K. Singh and C. A. Longhurst, Artificial intelligence note summarization in the emergency department, *JAMA Network Open* **9**, e2618021 (06 2026).
6. B. Edelman, H. Kim and J. Skolnick, Scalable discovery and validation of order-specific electronic health record event trajectories for interpretable adverse-outcome risk estimation, *Journal of Biomedical Informatics* **177**, p. 105008 (2026).
7. Y. Kim, J. Lee, S. B. Koh and K. H. Lee, An integrated evaluation framework for synthetic clinical data in severely imbalanced settings: fidelity, privacy-risk profiling, and diagnostic utility, *BMC Medical Informatics and Decision Making* (2026).
8. G. Moharasar and T. B. Ho, A semi-supervised approach for temporal information extraction from clinical text, in *2016 IEEE RIVF international conference on computing & communication technologies, research, innovation, and vision for the future (RIVF)*, 2016.
9. E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann and M. B. A. McDermott, Publicly available clinical bert embeddings (2019).
10. R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon and T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, *Briefings in Bioinformatics* **23**, p. bbac409 (09 2022).
11. R. Chaturvedi, P. Baghershahi, S. Medya and B. Di Eugenio, Temporal relation extraction in clinical texts: A span-based graph transformer approach, *arXiv preprint arXiv:2503.18085* (2025).
12. M. A. H. Kabir, A. Almulhim, X. Luo and M. A. Hasan, Informative causality extraction from medical literature via dependency-tree-based patterns, *Journal of Healthcare Informatics Research* **6**, 295 (may 2022).
13. V. Sciannameo, D. J. Pagliari, S. Urru, P. Grimaldi, H. Ocagli, S. Ahsani-Nasab, R. I. Comoretto, D. Gregori and P. Berchiolla, Information extraction from medical case reports using openai instructgpt, *Computer Methods and Programs in Biomedicine* **255**, p. 108326 (2024).
14. Y. Zhou, W.-T. Chen, B. Zhang, D. Lee, J. H. Caufield, K.-W. Chang, Y. Sun, P. Ping and W. Wang, Create: Clinical report extraction and annotation technology (feb 2021).

15. Z. Zhao, Q. Jin, F. Chen, T. Peng and S. Yu, A large-scale dataset of patient summaries for retrieval-based clinical decision support systems, *Scientific Data* **10** (2022).
16. E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart and J. Sun, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism (2017).
17. Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi and G. Salimi-Khorshidi, Behrt: transformer for electronic health records, *Scientific reports* **10**, p. 7155 (2020).
18. L. Rasmy, Y. Xiang, Z. Xie and C. S. and Jianying Hu, Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, *npj Digital Medicine* **4**, p. 86 (jun 2021).
19. P. Jiang, C. Xiao, A. Cross and J. Sun, Graphcare: Enhancing healthcare predictions with personalized knowledge graphs, *arXiv preprint arXiv:2305.12788* (2023).
20. J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher and S. McLachlan, Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, *Journal of the American Medical Informatics Association* **25**, 230 (mar 2018).
21. J. Li, B. J. Cairns, J. Li, A. D. Gilden, M.-I. Z. Li and J. H. Saltz, Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications, *npj Digital Medicine* **6**, p. 98 (jun 2023).
22. J. Yoon, M. Mizrahi, N. F. Ghalaty, M. S. Oberst and P. Szolovits, Ehr-safe: generating high-fidelity and privacy-preserving synthetic electronic health records, *npj Digital Medicine* **6**, p. 141 (sep 2023).
23. B. Theodorou, C. Xiao and J. Sun, Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model, *Nature Communications* **14**, p. 5305 (sep 2023).
24. M. Falis, A. P. Gema, H. Dong, L. Daines, S. Basetti, M. Holder, R. S. Penfold, A. Birch and B. Alex, Can gpt-3.5 generate and code discharge summaries?, *Journal of the American Medical Informatics Association* **31**, p. 2284–2293 (oct 2024).
25. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi, Bertscore: Evaluating text generation with bert, *arXiv preprint arXiv:1904.09675* (2019).
26. P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics* **20**, 53 (1987).
27. T. Caliński and J. Harabasz, A dendrite method for cluster analysis, *Communications in Statistics-theory and Methods* **3**, 1 (1974).
28. S. Holm, A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **6**, 65 (1979).

MedSR: Aligning Language Models to Reason and Search for Medical Question Answering with Group Relative Policy Optimization

Zaifu Zhan^{1,2}, Jun Wang¹, Yiran Song¹, Min Zeng¹, Kai Yu¹, Mengyuan Cui¹,
Mingquan Lin¹, Feng Xie¹, and Rui Zhang^{1,*}

¹*Division of Computational Health Sciences, Department of Surgery, University of Minnesota, Minneapolis, MN, USA*

²*Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA*

**Corresponding author. E-mail: ruizhang@umn.edu*

Large language models can answer medical questions from parametric memory, but this behavior gives limited control over whether the model seeks evidence or uses retrieved information appropriately. We present MedSR, a medical search-and-reasoning framework that fine-tunes language models to produce auditable reason–search–answer trajectories with constrained final answers. MedSR uses group relative policy optimization with LoRA adaptation, an external search tool, and rule-based answer rewards, without a separately trained reward model. We evaluate three compact models on MedQA-USMLE, HeadQA, and PubMedQA, comparing direct prompting, chain-of-thought prompting, self-consistency, few-shot prompting, untrained search prompting, and GRPO-trained search behavior. Four retrieval corpora are studied: PubMed, medical textbooks, StatPearls, and Wikipedia. The results were task dependent: trained search gave the clearest gains for Qwen2.5-3B and Llama-3.2-3B on MedQA and HeadQA, while the PubMed corpus was most closely aligned with PubMedQA. In this benchmark setting, updating the model to generate and use search queries was more effective than simply appending retrieved passages to the prompt, and performance depended on the retrieval corpus, reward design, retrieval depth, and adapter capacity.

Keywords: Medical Question Answering; Retrieval-Augmented Generation; Reinforcement Learning; Group Relative Policy Optimization; Large Language Models.

1. Introduction

Since GPT-3¹ and its successors demonstrated that scaling Transformer-based models^{2,3} could produce surprisingly capable text generators, the NLP community has pushed these models into increasingly specialized domains. Medicine stands out among them, not because question answering is new, but because the cost of a wrong answer is unusually high. General-purpose and domain-adapted models now perform well on clinical knowledge benchmarks and medical examination questions.^{4–9} Recent medical foundation-model studies have also emphasized efficient adaptation, quantization, dataset composition, retrieval-based conditioning, and reliability in medical reasoning.^{10–14} Yet answers still often come from the model’s parametric

memory alone, a limitation in a field where evidence shifts with new studies, guidelines, and drug approvals.

Retrieval-augmented models such as REALM,¹⁵ DPR,¹⁶ and RAG¹⁷ were developed to reduce this dependence on memorized knowledge by pairing generation with external documents. Later systems showed that retrieval can support both open-domain question answering and few-shot language modeling when the retrieved passages are well matched to the input.^{18–20} In medical QA, the documents may come from PubMed abstracts, textbooks, clinical summaries, or broad reference sources. Corpus choice and retrieval depth matter more than early RAG applications often assumed, and recent biomedical in-context learning work has also shown that retrieval strategy affects downstream performance.^{13,21,22} The retrieval behavior, however, is usually fixed before the model sees a question. A developer chooses the query template, corpus, top- k , and truncation length. That recipe may be adequate for some questions and wasteful for others. It can also retrieve text that matches the surface wording of a long question while missing the fact needed to answer it.

We turned to reinforcement learning for a practical reason: medical QA already has a usable reward signal. Benchmarks such as MedQA provide a ground-truth answer, so a full trajectory—reasoning, search, evidence reading, and final answer—can be scored without training a separate reward model. Policy-gradient methods such as PPO established a practical route for optimizing generated behavior beyond next-token likelihood,²³ and GRPO removes the need for a learned critic in settings with group-level rewards.²⁴ Recent work has also shown that models can interleave reasoning with browsing, tool use, and search.^{25–29} At the same time, medical LLM reviews and self-correction studies suggest that explicit reasoning traces do not by themselves guarantee correct medical answers.^{14,30,31} Current medical RAG pipelines rarely make that connection. Retrieval is still usually a preprocessing step, and the learning objective seldom treats the query, evidence, and final medical answer as one trajectory.

MedSR studies this formulation in medical QA. It fine-tunes compact instruction-tuned language models to learn what to search and how to use retrieved evidence before committing to a constrained answer. MedSR uses GRPO with LoRA adaptation, a medical search tool, and rule-based answer rewards to optimize the reason–search–answer trajectory. We evaluate MedSR on MedQA-USMLE,³² HeadQA,³³ and PubMedQA,³⁴ across three base models and four retrieval corpora. These benchmarks support controlled comparison, but they are narrower than real clinical language tasks; recent holistic evaluations such as MedHELM emphasize that medical LLM assessment should span more diverse task types and output formats.³⁵ We compare trained retrieval with direct prompting, chain-of-thought, self-consistency, few-shot prompting, and untrained search prompting.

2. Materials and Methods

2.1. Framework overview

Fig. 1 shows how a question passes through MedSR. The policy is the language model whose search-and-answer behavior we optimize. Its pretrained weights remain frozen while LoRA adapters are updated. Each experiment connects the model to one of the four retrieval corpora shown in the search-tool block. During a rollout, the model generates a query, receives top- k

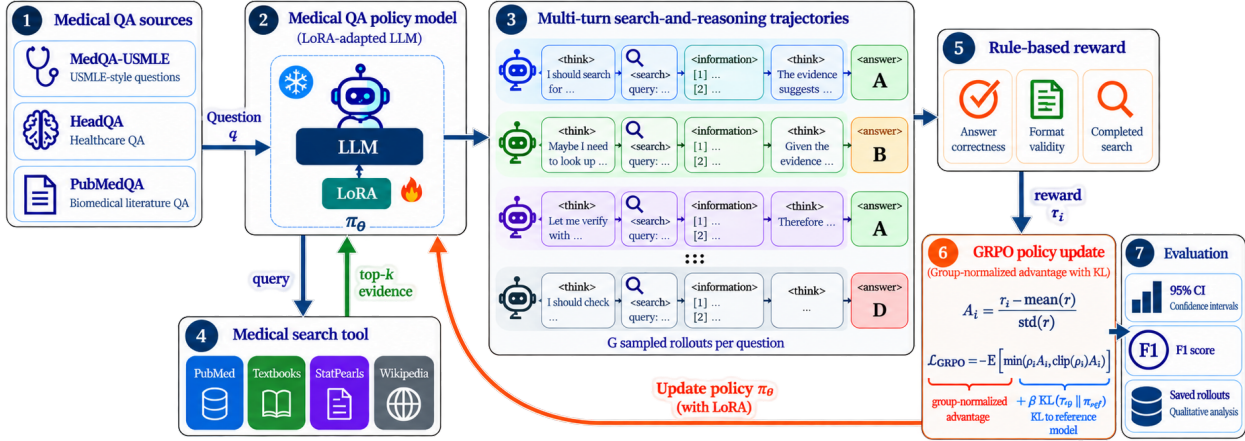


Fig. 1. **MedSR architecture for learning medical search and reasoning with GRPO.** Questions from three datasets (1) enter a LoRA-adapted model (2), which samples multiple trajectories per question (3) using top- k evidence from one corpus per experiment (4). Answer correctness and trajectory-format checks determine rewards (5); completed search is part of the format check. GRPO updates the adapters under KL regularization while base weights remain frozen (6). Evaluation reports answer metrics and retains rollout traces (7).

snippets from that corpus, and continues reasoning before producing an answer in the required format.

Inspired by Search-R1,²⁶ MedSR treats search as an action inside the RL trajectory and redesigns the setting for medical QA. We replace general web search with medical evidence corpora, constrain the final answer to the benchmark label space, and score the completed trajectory with medical QA rewards. MedQA-USMLE and HeadQA require a single option letter, while PubMedQA requires *yes*, *no*, or *maybe*. We also store the generated query, retrieved snippets, reasoning text, and final answer for error analysis. The reward is assigned only after the answer is generated, so the model must learn which query and evidence-use pattern lead to a correct medical answer.

During training, the same model samples multiple trajectories for each question. Completed trajectories receive rewards based on answer correctness and format validity. GRPO compares rewards within the group and updates the LoRA adapters using the resulting advantages, with KL regularization against a frozen reference model. Supervision comes from benchmark answers rather than manually annotated search traces. The evaluation block summarizes answer-level metrics and saved trajectories for qualitative analysis.

Our contribution is the medical QA implementation and evaluation of this search-as-action approach: dataset-specific answer constraints, parameter-efficient adaptation, and an analysis of evidence sources and failure modes across three models, three datasets, and four corpora.

2.2. Datasets

We use MedQA-USMLE,³² HeadQA,³³ and PubMedQA³⁴ labeled splits. HeadQA covers multiple healthcare disciplines, including medicine, nursing, and pharmacology. We split PubMedQA labeled into 500 training examples and 500 test examples. We treat MedQA and HeadQA as multiple-choice tasks requiring a single option letter, whereas PubMedQA requires *yes*, *no*, or *maybe*. We convert each example into a multiturn tool-use record containing the prompt, ground-truth answer list, original question, default search query, and search-tool metadata; for PubMedQA, we append abstract context to the question text before prompting.

2.3. Prompt and tool-use trajectory

We define the learning unit as a structured tool-use trajectory. Following the search-as-action design used in Search-R1,²⁶ the policy uses `<think>` for reasoning, `<search>` for the query, `<information>` for retrieved evidence, and `<answer>` for the final response. MedQA and HeadQA require a single option letter; PubMedQA requires *yes*, *no*, or *maybe*. With this format, search and evidence remain auditable, and the same reward function can score answer correctness and minimal trajectory validity across datasets.

2.4. Retrieval corpora

We evaluated four MedRAG-style corpora: PubMed, medical textbooks, StatPearls, and Wikipedia.²¹ They contained 23,898,701, 125,847, 365,650, and 1,179,078 snippets, respectively, spanning article-level evidence, educational exposition, clinical summaries, and broad background knowledge. Dense retrieval used E5-base-v2 embeddings³⁶ and FAISS Flat indexes.³⁷ Queries and passages used E5 prefixes, embeddings were L2-normalized, and the retrieval service returned ranked snippets with title/identifier, score when available, and truncated text. Main GRPO experiments used top- $k = 2$ with a 900-character cap per document to limit rollout context. Search-prompt baselines retained the baseline script’s top- $k = 3$ setting with the same per-document cap. These comparisons therefore reflect both fine-tuning and retrieval-depth settings; the top- k ablation measures sensitivity within the trained system.

2.5. GRPO fine-tuning

We use GRPO as a critic-free reinforcement learning objective.²⁴ For each prompt x , the rollout engine samples a group of $G = 4$ complete tool-use trajectories $\{y_i\}_{i=1}^G$, where each trajectory includes generated reasoning tokens, search actions, retrieved observations, and the final answer. We compute the scalar reward r_i only after the full trajectory is completed. Instead of training a value function, GRPO estimates the advantage by normalizing rewards within the group:

$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G) + \epsilon}.$$

The normalization compares trajectories sampled for the same question rather than absolute reward scales across datasets.

We update the actor with a clipped policy-gradient objective over generated tokens, with a KL regularizer to keep the LoRA-adapted policy close to the frozen reference model:

$$\mathcal{L}_{\text{GRPO}} = -\frac{1}{G} \sum_{i=1}^G \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \min(\rho_{i,t} A_i, \text{clip}(\rho_{i,t}, 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}}) A_i) + \beta D_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{ref}}),$$

where $\rho_{i,t} = \pi_{\theta}(y_{i,t} \mid x, y_{i,<t}) / \pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})$, $\beta = 0.001$, and π_{ref} is the frozen base model. We apply the loss only to generated assistant tokens and treat retrieved text as an external observation. Main experiments used training splits for optimization and test splits for evaluation; ablations varied retrieval depth, reward design, corpus/retriever condition, LoRA rank, and training duration in the Llama-3.2-3B-Instruct PubMedQA setting.

Implementation used the verl GRPO trainer. We adapted the actor with LoRA on all linear modules, using rank 64 and alpha 32 in the main experiments.³⁸ We used the frozen reference model for KL regularization. Training used bfloat16 precision, gradient checkpointing, remove-padding optimization, FSDP parameter and optimizer offload, and vLLM rollout generation.³⁹ The learning rate was 3×10^{-5} ; train, mini-batch, and validation batch sizes were 4; each prompt produced 4 rollouts; and maximum prompt, response, and tool-response lengths were 2048, 512, and 1024 tokens, respectively.

2.6. Reward function

We extract the final answer from the last `<answer>...</answer>` span and evaluate only that answer, not the reasoning text. Responses receive zero answer reward if the answer is absent or outside the dataset-specific answer space. Valid answers are lowercased and stripped of punctuation and whitespace before exact matching.

The main experiments use a shaped reward that combines exact answer accuracy with a small trajectory-format term:

$$r = \begin{cases} 1.0, & \text{if the answer is correct and the format is valid,} \\ 0.9, & \text{if the answer is correct but the format is invalid,} \\ 0.1, & \text{if the answer is incorrect but the format is valid,} \\ 0, & \text{otherwise.} \end{cases}$$

The format criterion required a valid answer, at least one search span, at least one information span, and balanced trajectory tags. Thus, “Completed search” in Fig. 1 refers to the search/information tag check within the format term, not a separate main-experiment reward or a judgment of evidence relevance. We keep this term small so that reward primarily reflects answer correctness. Table 1 summarizes ablated reward variants.

3. Experimental Setup

3.1. Baselines

We run baselines with vLLM separately from the GRPO trainer.³⁹ All baselines are inference-only prompting methods; none are trained with GRPO. Direct answering asks the model to return the constrained answer without retrieved evidence. Chain-of-thought prompting

Table 1. **Reward formulations for GRPO fine-tuning of medical question-answering models.** Answer + format is the main-experiment reward; the other formulations vary the contributions of answer correctness, trajectory format, search use, or response length. All apply the same dataset-specific answer-validity check before scoring.

Variant	Reward definition
Answer only	$r = 1$ if correct; 0 otherwise
Answer + format	$r = 0.9$ if correct, plus 0.1 if format is valid
Format-heavy	$r = 0.8$ if correct, plus 0.2 if format is valid
Search-light	$r = 0.85$ if correct, plus 0.1 if format is valid, plus 0.05 if search is used
Length penalty	$r = 0.9$ if correct, plus 0.1 if format is valid, minus 0.02 if over length

allows intermediate reasoning before the final answer,⁴⁰ and self-consistency samples 5 chain-of-thought completions at temperature 0.7 and uses majority vote.⁴¹ Few-shot prompting provides one, three, five, or ten labeled examples in the prompt. Untrained search prompting appends retrieved snippets to the prompt but does not update model weights. Direct and few-shot baselines use a 64-token cap; chain-of-thought and search baselines use 512 tokens. We rescore all outputs with the same answer extractor and normalization rules.

3.2. Evaluation

We use exact-match accuracy after answer extraction as the primary metric, with Wilson 95% confidence intervals.⁴² Micro-F1 pools decisions across labels and equals accuracy in this single-label evaluation; macro-F1 averages F1 equally over observed reference and predicted labels, while weighted-F1 weights each label by its reference frequency. Empty extracted answers are assigned an invalid label for F1 calculation. The format score checks search-trajectory structure as defined above. Baselines are scored through an answer-only representation, so their zero format scores do not imply invalid answer labels or failure to follow the baseline prompt. Shaped reward combines accuracy and format; it is not a clinical-quality score. We compute metrics from saved per-example outputs. Code for the MedSR experiments is available at <https://github.com/Learner4everrr/MedSR.git>.

4. Results

4.1. Main comparison across models, datasets, and corpora

Table 2 presents the central comparison by benchmark, method, corpus, and model, reporting accuracy, Wilson 95% confidence intervals, F1 scores, and format score. The relative ranking of methods varied across datasets and models, so we focus on trends that were consistent across multiple settings. Direct answering, chain-of-thought, self-consistency, and few-shot prompting remained competitive for several model–dataset pairs, particularly when the answer format was simple or the base model already appeared to know the relevant fact. Longer reasoning alone did not consistently improve accuracy. Chain-of-thought and self-consistency sometimes helped, but they also produced verbose generations that still ended with the wrong option or PubMedQA label.

Untrained search prompting did not consistently exceed the best prompt-only baseline, even for PubMedQA. Review of the saved generations suggested that retrieved text sometimes introduced a related but answer-irrelevant finding or increased format errors. MedSR fine-tuning performed better when the model learned to issue the query, read the returned snippet, and keep the final answer inside the required answer space.

The strongest gains appeared for Qwen2.5-3B and Llama-3.2-3B-Instruct on MedQA and HeadQA. DeepSeek-R1-Distill-Qwen-1.5B changed less, which is consistent with the qualitative rollouts: this smaller reasoning-distilled model was less stable in following the tool-use format. Corpus choice changed the ranking as well. PubMed aligned best with PubMedQA, while textbooks, StatPearls, and Wikipedia were more useful for some examination-style questions. Detailed model-specific values are reported in Table 2, and cross-condition trends are shown in Fig. 2.

4.2. *Qualitative success and failure cases*

We reviewed saved test-set generations from trained-search experiments. Table 3 shows cases stratified by retrieval quality and final-answer correctness. In successful cases, the generated query often compressed a long medical question into a disease, intervention, mechanism, or study finding that matched the retrieved evidence. The model then used the evidence to select a constrained answer rather than copying a sentence from the retrieved passage.

The failure cases were heterogeneous. Some rollouts retrieved relevant passages but selected an incorrect option, indicating that evidence access did not guarantee correct evidence interpretation. Other rollouts retrieved broad or off-topic snippets, leaving the final answer primarily dependent on the model’s parametric memory. A third class involved trajectory or answer-format errors, such as placing explanatory text inside the answer tag or returning evidence-like phrases instead of a single option letter or PubMedQA label. Looking at the saved trajectory made these errors separable: a reviewer could see whether the failure began with the query, the retrieved text, the interpretation of relevant evidence, or the final answer format.

4.3. *Error analysis*

The saved rollouts made the error analysis more specific than aggregate accuracy. Several failures started at the query step. The model sometimes searched with an overly broad disease name, copied too much of the question, or omitted the key intervention, population, or outcome; the retrieved evidence then looked plausible but did not resolve the question. A different failure came from corpus mismatch. Literature-style snippets were not always optimal for examination questions that required canonical teaching facts, while educational summaries could miss PubMedQA-specific study details.

Other errors happened after apparently useful retrieval. The model could find topically relevant text but map it to the wrong option or yes/no/maybe label, especially when the evidence contained negation, comparative wording, or a subtle distinction between association and causation. In a smaller set of cases, the final answer appeared to come from parametric memory even when retrieved evidence suggested another label. We also saw plain format

Table 2. **Performance of MedSR on MedQA-USMLE, HeadQA, and PubMedQA.** Search uses retrieved context without weight updates; Ours denotes MedSR. Acc: accuracy; CI: Wilson 95% confidence interval; Micro, Macro, Wt: micro-, macro-, and weighted-F1; Fmt: search-trajectory format score. Metrics are percentages. Bold identifies the highest values within each dataset-model block. Text, StatP, and Wiki denote textbooks, StatPearls, and Wikipedia.

		Qwen2.5-3B						Llama-3.2-3B-Instruct						DeepSeek-R1-Distill-Qwen-1.5B					
Experiment	Corpus	Acc	95% CI	Micro	Macro	Wt	Fmt	Acc	95% CI	Micro	Macro	Wt	Fmt	Acc	95% CI	Micro	Macro	Wt	Fmt
MedQA																			
Direct	-	32.3	29.8–34.9	32.3	30.6	38.6	0.0	51.2	48.5–54.0	51.2	40.4	50.9	0.0	23.6	21.4–26.1	23.6	16.9	25.7	0.0
CoT	-	31.0	28.5–33.6	31.0	24.9	37.5	0.0	48.9	46.1–51.6	48.9	34.2	51.5	0.0	14.2	12.4–16.2	14.2	12.8	19.2	0.0
SC-CoT	-	44.2	41.5–47.0	44.2	29.4	44.6	0.0	47.7	44.9–50.4	47.7	40.1	50.5	0.0	24.5	22.2–26.9	24.5	17.0	25.6	0.0
1-shot	-	41.6	39.0–44.4	41.6	33.9	43.0	0.0	50.5	47.8–53.3	50.5	40.6	51.2	0.0	16.0	14.1–18.1	16.0	10.5	21.2	0.0
3-shot	-	43.2	40.5–45.9	43.2	35.2	44.7	0.0	51.0	48.2–53.7	51.0	40.5	51.0	0.0	13.4	11.7–15.4	13.4	14.1	17.6	0.0
5-shot	-	39.2	36.6–41.9	39.2	33.2	42.0	0.0	48.1	45.3–50.8	48.1	38.5	48.5	0.0	12.1	10.4–14.0	12.1	9.1	16.0	0.0
10-shot	-	31.2	28.7–33.8	31.2	24.8	37.7	0.0	41.5	38.8–44.2	41.5	29.7	37.6	0.0	14.5	12.6–16.5	14.5	10.7	18.8	0.0
Search	PubMed	29.7	27.2–32.3	29.7	28.0	35.5	0.0	47.0	44.2–49.7	47.0	31.1	46.9	0.0	11.3	9.7–13.2	11.3	8.8	15.6	0.0
Search	Text	30.3	27.9–32.9	30.3	28.6	36.2	0.0	48.6	45.9–51.4	48.6	32.3	48.6	0.0	12.6	10.9–14.6	12.6	11.7	17.6	0.0
Search	StatP	32.5	30.0–35.1	32.5	29.8	37.8	0.0	48.5	45.8–51.3	48.5	32.2	48.3	0.0	13.0	11.2–14.9	13.0	10.4	18.2	0.0
Search	Wiki	32.8	30.3–35.5	32.8	30.2	37.9	0.0	45.0	42.3–47.8	45.0	30.6	45.9	0.0	12.5	10.8–14.4	12.5	13.6	16.9	0.0
Ours	PubMed	49.9	47.1–52.6	49.9	8.2	49.6	0.1	54.2	51.5–56.9	54.2	44.3	55.7	95.1	24.7	22.5–27.2	24.7	14.0	24.9	0.0
Ours	Text	49.5	46.7–52.2	49.5	38.6	49.0	98.7	56.2	53.5–58.9	56.2	56.1	56.3	100	26.5	24.1–29.0	26.5	7.4	26.6	0.0
Ours	StatP	48.4	45.7–51.1	48.4	36.7	46.8	99.5	54.7	51.9–57.4	54.7	54.6	54.8	100	26.8	24.4–29.3	26.8	9.4	12.9	95.8
Ours	Wiki	51.3	48.6–54.0	51.3	40.8	51.2	99.6	53.8	51.1–56.5	53.8	53.3	53.7	99.3	25.8	23.5–28.3	25.8	21.0	26.8	0.0
HeadQA																			
Direct	-	47.4	45.6–49.3	47.4	40.7	51.3	0.0	56.9	55.0–58.7	56.9	45.1	56.6	0.0	36.4	34.6–38.2	36.4	21.4	37.4	0.0
CoT	-	50.9	49.0–52.8	50.9	36.1	54.4	0.0	55.8	54.0–57.7	55.8	38.4	57.7	0.0	32.3	30.6–34.1	32.3	20.0	35.0	0.0
SC-CoT	-	57.0	55.2–58.9	57.0	37.8	57.0	0.0	56.3	54.5–58.2	56.3	45.9	57.4	0.0	37.4	35.6–39.2	37.4	30.0	37.5	0.0
1-shot	-	45.2	43.3–47.1	45.2	33.7	42.9	0.0	52.8	50.9–54.6	52.8	35.8	53.7	0.0	31.1	29.4–32.8	31.1	13.4	33.6	0.0
3-shot	-	49.5	47.7–51.4	49.5	40.7	51.5	0.0	55.0	53.1–56.8	55.0	54.6	54.6	0.0	31.4	29.7–33.1	31.4	22.8	34.2	0.0
5-shot	-	51.3	49.4–53.1	51.3	41.9	53.0	0.0	54.1	52.2–55.9	54.1	53.8	53.9	0.0	28.7	27.0–30.4	28.7	18.3	32.1	0.0
10-shot	-	51.5	49.7–53.4	51.5	43.1	54.4	0.0	52.6	50.7–54.4	52.6	42.1	52.7	0.0	28.3	26.6–30.0	28.3	16.0	32.1	0.0
Search	PubMed	33.0	31.3–34.8	33.0	33.1	41.5	0.0	56.8	55.0–58.7	56.8	38.0	57.0	0.0	33.7	31.9–35.5	33.7	25.4	38.1	0.0
Search	Text	35.4	33.6–37.2	35.4	34.6	43.3	0.0	55.7	53.9–57.6	55.7	37.2	55.8	0.0	32.6	30.8–34.3	32.6	24.6	37.0	0.0
Search	StatP	35.4	33.7–37.3	35.4	34.4	43.0	0.0	53.5	51.6–55.4	53.5	35.6	53.5	0.0	33.0	31.2–34.8	33.0	29.6	37.0	0.0
Search	Wiki	35.4	33.7–37.3	35.4	34.1	42.7	0.0	53.4	51.5–55.3	53.4	35.9	53.9	0.0	30.8	29.1–32.5	30.8	19.8	34.7	0.0
Ours	PubMed	63.3	61.5–65.1	63.3	50.9	63.7	100	62.4	60.5–64.2	62.4	62.3	62.4	99.8	37.2	35.4–39.0	37.2	25.2	37.8	0.0
Ours	Text	63.4	61.6–65.2	63.4	41.9	63.0	99.7	62.5	60.6–64.3	62.5	31.2	62.5	99.6	35.9	34.1–37.7	35.9	13.3	36.8	0.0
Ours	StatP	65.4	63.6–67.2	65.4	64.9	65.1	100	62.9	61.1–64.7	62.9	62.8	62.9	99.9	36.6	34.8–38.4	36.6	8.7	37.2	0.0
Ours	Wiki	65.5	63.7–67.2	65.5	52.0	65.2	100	61.8	59.9–63.6	61.8	41.5	62.3	98.1	37.7	36.0–39.6	37.7	17.1	38.5	0.0
PubMedQA																			
Direct	-	60.4	56.0–64.6	60.4	31.1	57.8	0.0	69.6	65.4–73.5	69.6	36.8	67.2	0.0	60.0	55.6–64.2	60.0	35.8	61.6	0.0
CoT	-	59.4	55.0–63.6	59.4	32.0	59.0	0.0	70.4	66.3–74.2	70.4	36.9	67.2	0.0	31.2	27.3–35.4	31.2	23.9	40.6	0.0
SC-CoT	-	54.4	50.0–58.7	54.4	41.0	53.8	0.0	69.4	65.2–73.3	69.4	48.3	66.3	0.0	65.8	61.5–69.8	65.8	38.1	64.9	0.0
1-shot	-	65.2	60.9–69.2	65.2	34.7	62.4	0.0	68.0	63.8–71.9	68.0	47.1	64.5	0.0	53.6	49.2–57.9	53.6	31.8	56.4	0.0
3-shot	-	57.8	53.4–62.1	57.8	30.8	55.6	0.0	52.0	47.6–56.3	52.0	25.8	50.3	0.0	49.6	45.2–54.0	49.6	30.3	51.4	0.0
5-shot	-	67.2	63.0–71.2	67.2	35.1	64.2	0.0	64.8	60.5–68.9	64.8	42.1	59.6	0.0	51.0	46.6–55.4	51.0	30.2	53.8	0.0
10-shot	-	67.0	62.8–71.0	67.0	34.1	63.2	0.0	56.4	52.0–60.7	56.4	36.6	52.3	0.0	48.0	43.7–52.4	48.0	29.4	51.8	0.0
Search	PubMed	62.6	58.3–66.7	62.6	34.2	60.9	0.0	71.6	67.5–75.4	71.6	49.7	68.0	0.0	41.0	36.8–45.4	41.0	24.3	46.3	0.0
Search	Text	41.4	37.2–45.8	41.4	23.2	38.3	0.0	64.0	59.7–68.1	64.0	45.4	61.7	0.0	26.0	22.3–30.0	26.0	19.4	35.9	0.0
Search	StatP	45.0	40.7–49.4	45.0	28.0	44.5	0.0	64.8	60.5–68.9	64.8	45.5	62.3	0.0	26.6	22.9–30.6	26.6	20.0	36.6	0.0
Search	Wiki	35.4	31.3–39.7	35.4	20.2	30.4	0.0	63.6	59.3–67.7	63.6	45.8	61.3	0.0	33.8	29.8–38.1	33.8	24.4	44.3	0.0
Ours	PubMed	73.4	69.4–77.1	73.4	38.3	69.8	99.8	74.6	70.6–78.2	74.6	51.6	70.5	100	70.6	66.5–74.4	70.6	48.2	66.4	0.0
Ours	Text	70.8	66.7–74.6	70.8	24.7	67.6	99.4	64.2	59.9–68.3	64.2	45.1	61.1	99.8	67.0	62.8–71.0	67.0	34.1	63.0	0.0
Ours	StatP	73.4	69.4–77.1	73.4	38.2	69.7	99.8	70.8	66.7–74.6	70.8	36.4	66.7	99.8	70.6	66.5–74.4	70.6	36.2	66.7	0.0
Ours	Wiki	68.0	63.8–71.9	68.0	28.5	64.9	99.6	73.0	68.9–76.7	73.0	38.2	69.3	99.8	68.6	64.4–72.5	68.6	46.5	64.4	0.0

failures: explanatory text, retrieved phrases, or multiple candidate labels were placed inside `<answer>...</answer>` where only one constrained answer was allowed.

The error analysis separates the behaviors improved by MedSR from the behaviors that remain weakly supervised. GRPO fine-tuning can reward trajectories that search and end with valid answers, but the reward only observes final-answer correctness and coarse format validity. It does not directly supervise whether the query is medically optimal, whether each retrieved passage is clinically relevant, or whether the reasoning faithfully follows the evidence. Future reward designs should consider retrieval quality, evidence attribution, contradiction handling,

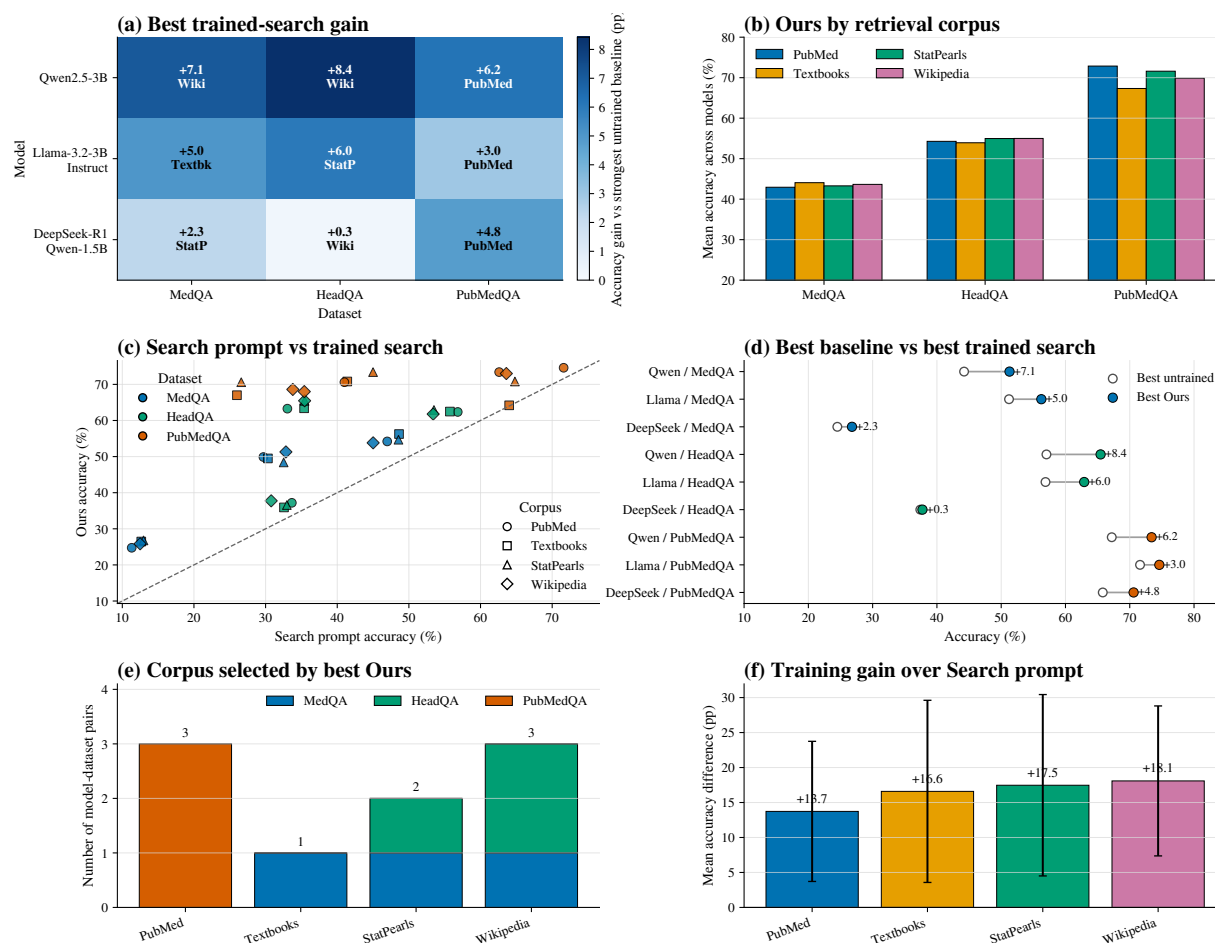


Fig. 2. **Comparison of MedSR and inference-only baselines across models, datasets, and retrieval corpora.** (a) Best MedSR minus best baseline accuracy per model–dataset pair. (b) MedSR accuracy averaged across models for each dataset and corpus. (c) Search versus MedSR accuracy, paired by model, dataset, and corpus. (d) Best baseline and MedSR accuracies. (e) Number of model–dataset pairs for which each corpus yields the highest MedSR accuracy; stacked colors denote datasets. (f) MedSR minus Search accuracy by corpus, averaged over nine model–dataset pairs; error bars show one standard deviation across pairs. Accuracy differences are in percentage points. Best values denote observed test-set maxima.

and clinician-rated utility.

4.4. Ablation studies

The ablations use the Llama-3.2-3B-Instruct PubMedQA setting and vary retrieval depth, reward formulation, corpus/retriever condition, LoRA rank, and fine-tuning duration. These experiments were designed to separate three contributors to performance: the quality of evidence exposed to the policy, the reward signal used to select higher-quality trajectories, and the adapter capacity available for changing search-and-answer behavior.

Retrieval depth was non-monotonic in Fig. 3(a). Top- $k = 2$ was strongest, while larger k values often reduced accuracy. This pattern is relevant for medical QA because additional

Table 3. **Retrieved evidence and answer outcomes in selected MedSR test-set trajectories.** Qwen2.5-3B examples illustrate correct answers with aligned evidence, answer or format errors despite aligned retrieval, and retrieval failures. Questions and retrieved passages are excerpted; reference and generated answers are shown for each case.

Dataset / corpus	Retrieval quality	Example question	Top retrieved document/snippet shown to the model	Target / model answer	Outcome
PubMedQA / PubMed	Aligned	Bone-anchor fixation for percutaneous bladder neck suspension in female stress urinary incontinence.	PubMed hit 0322_22353: “Does bone anchor fixation improve the outcome of percutaneous bladder neck suspension in female stress urinary incontinence?”	yes / yes	Correct answer after aligned retrieval.
HeadQA / StatPearls	Aligned	“Form extracellular fibers with high tensile strength.”	StatPearls article 19708_0: “Biochemistry, Collagen Synthesis – Introduction. Collagen ... provides structural support ... resistance to stretching.”	B / B	Correct option after aligned retrieval.
HeadQA / Textbooks	Aligned	“Fibrocartilage.”	Histology Ross: “Fibrocartilage is a combination of dense connective tissue and cartilage ... large bundles of type I collagen.”	C / C	Correct option after aligned retrieval.
PubMedQA / PubMed	Aligned	Optimal acid suppression time for healing severe reflux oesophagitis.	PubMed hit 0661_16389: “A model of healing of Los Angeles grades C and D reflux oesophagitis: is there an optimal time of acid suppression...”	yes / span	Retrieval aligned, but final answer violated the yes/no/maybe format.
PubMedQA / PubMed	Aligned	Whether a practicing surgeon can detect early lymphedema reliably.	PubMed hit 0477_21998: “Can a practicing surgeon detect early lymphedema reliably? ... circumference changes ... limb volume.”	maybe / no	Retrieval aligned, but the model made a semantic yes/no/maybe error.
HeadQA / PubMed	Failed/empty	“The linear molecule of DNA associated with proteins is.”	Query: “...”; PubMed hits 0975_2954 and 0976_10335: contents “. .”.	C / D	Retrieval failed; final answer was incorrect despite valid format.
MedQA-USMLE / Wikipedia	Off-topic/generic	Renal biopsy with intravascular spindle-shaped vacuoles after unstable angina treatment.	Wikipedia hit: “Urinary tract infection. Children”; second hit: “Kidney stone disease. Ureteroscopy surgery.”	B / D	Retrieval was off topic and final answer was incorrect.
MedQA-USMLE / Wikipedia	Off-topic/generic	Cocaine-associated chest pain after acute coronary syndrome treatment.	Wikipedia hit: “Heart. Imaging”; second hit: “Heart. Heart failure.”	D / A	Retrieval was off topic and final answer was incorrect.
HeadQA / Textbooks	Failed/incom	“The linear molecule of DNA associated with proteins is.”	Biochemistry Lippincott: “Linear and circular DNA molecules”; Cell Biology Alberts: “Polycomb group ... proteins.”	C / B	Retrieval was incomplete and final answer was incorrect.

snippets can add corroborating evidence, but they can also introduce partially related findings, population differences, or terminology that distracts the model from the required label. The corpus and retriever comparison in Fig. 3(c) showed a similar dependence on evidence alignment: PubMed was strongest for PubMedQA, whereas other corpora were competitive in exam-style settings. MedSR therefore depends on evidence source and retrieval depth, beyond the binary choice of using retrieval.

In the reward comparison in Fig. 3(b), answer correctness remained the most reliable training signal. Small format and search-shaping terms helped stabilize valid trajectories, but heavier format emphasis did not reliably improve accuracy. We use tags to make trajectories

auditable, but keep the reward focused on the final medical answer. In Fig. 3(d), rank 64 was the largest rank tested and performed best. This does not establish adapter-capacity saturation; higher ranks may further improve performance.

The checkpoint analysis in Fig. 3(e) showed that most improvement appeared within the first 500–1250 steps, followed by noisier later checkpoints. One possible explanation is early reinforcement of valid search-and-answer behavior, although accuracy alone cannot establish this mechanism. These curves are retrospective test-set sensitivity analyses. Repeated examination of test performance introduces a risk of selection bias; checkpoint and hyperparameter selection should use a separate validation split before final testing.

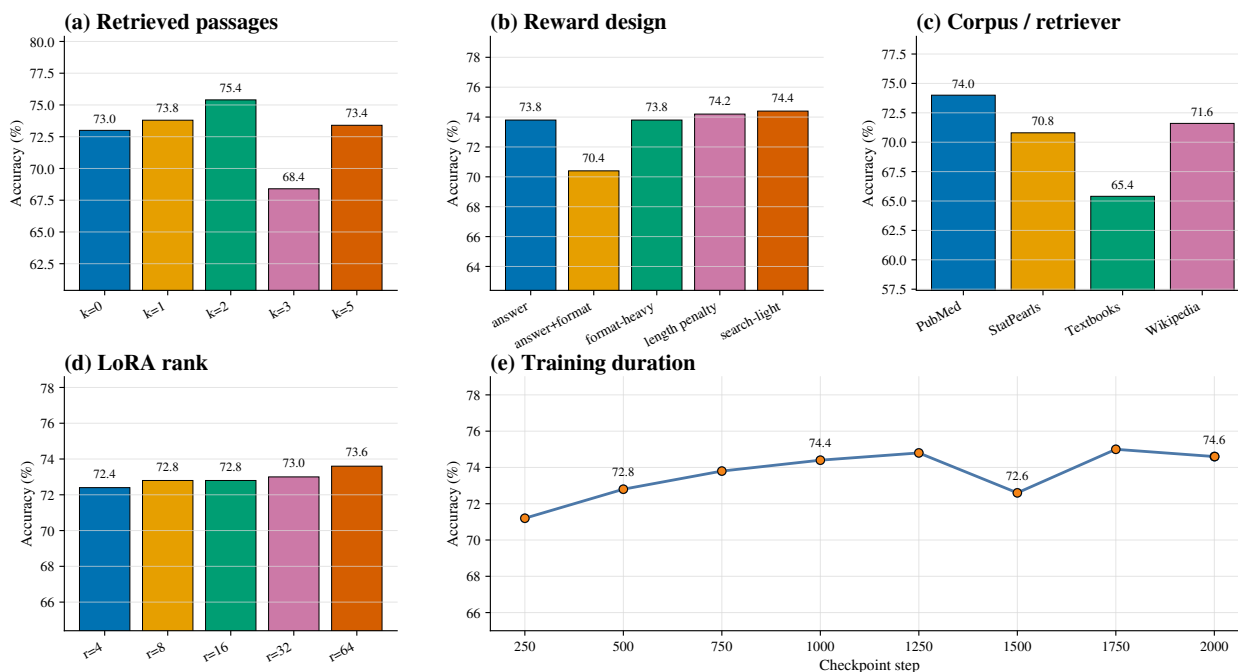


Fig. 3. Sensitivity of MedSR to retrieval configuration, reward formulation, adapter rank, and fine-tuning duration. All comparisons use Llama-3.2-3B-Instruct on PubMedQA. (a) Passages retrieved per query; $k = 0$ disables retrieval. (b) Reward formulations defined in Table 1. (c) Corpus choice with the E5-base-v2 retriever fixed. (d) LoRA adapter rank. (e) Test accuracy across fine-tuning checkpoints. All accuracies are percentages.

5. Discussion

Untrained search prompting was a relevant baseline for PubMedQA because the task is tied to abstracts, but adding retrieved text often left the model with the same answer-selection problem and, in some cases, extra distractors. Even snippets that are not copied may influence an answer through a plausible but irrelevant association or a population mismatch. MedSR uses GRPO-style relative rewards²⁴ to favor higher-scoring trajectories, although correct final answers alone do not establish faithful use of the evidence.

The corpus results also show that no single retrieval source was best across all benchmarks.

PubMed aligned best with PubMedQA, while textbooks, StatPearls, and Wikipedia were useful for different examination-style questions. Increasing top- k did not monotonically help either. Corpus and depth determine the evidence distribution that GRPO can learn from. A unified index spanning all four corpora would test access to mixed evidence sources; explicit source selection would additionally require a source-aware search action.

For medical informatics, the retrieval trace helps locate failures that final-answer accuracy cannot distinguish. The qualitative cases include relevant retrieval followed by an incorrect answer, sometimes involving negation, comparative wording, or association versus causation. These observations motivate evidence-attribution and contradiction checks, but do not establish the frequency of each failure type.

One practical use of MedSR is post-hoc audit. For an incorrect PubMedQA prediction, the saved trajectory helps reviewers examine the query, retrieved text, and evidence interpretation. A format checker followed by constrained regeneration could also reduce malformed answers. Such corrections should be evaluated separately from the original model outputs because they change the inference procedure.

The results also show that a model score should be interpreted as a system-level measurement rather than only as a property of the base LLM. For example, the same Llama-3.2-3B-Instruct model scored 45.0% with untrained Wikipedia search prompting and 56.2% with MedSR using textbooks on MedQA. A medical QA system built for literature questions may need PubMed-like evidence, while an examination-focused system may align better with educational or summary resources. More retrieval is not automatically better if additional passages add distractors or shift the model away from the answer format.

The study has limits that follow directly from this experimental design. We train and evaluate on structured QA benchmarks rather than live clinical decision support tasks, and final answers are constrained to benchmark-specific answer spaces. Because the rewards are based on benchmark labels, MedSR does not evaluate whether the model can communicate uncertainty, adapt to patient-specific context, or avoid harmful open-ended clinical advice. The experiments also use compact open models; larger models may follow the tool format more reliably, rely differently on parametric memory, or require different retrieval depths. Retrieved snippets should not be read as clinical guidelines or as sufficient evidence for care decisions. The retriever also remains fixed: MedSR trains the language model’s search-and-answer behavior, not the dense encoder or corpus index. Deployment would require local validation of corpus choice and retrieval depth, a separate validation split for checkpoint selection, safety evaluation, and clinician review.

6. Conclusion

We introduced MedSR, a reinforcement learning framework for medical question answering that trains language models to reason with external medical search. Across three benchmarks, three compact models, and four retrieval corpora, MedSR improved many model–dataset combinations and exposed an auditable reason–search–answer trace. In these experiments, appending retrieved passages to a frozen LLM was less effective than training the model to decide what to search for and how to use the returned evidence. For medical QA, search should

be evaluated as part of the learned answering process rather than only as an inference-time prompt addition.

References

1. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, Language models are few-shot learners, in *Advances in Neural Information Processing Systems*, eds. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin (Curran Associates, Inc., 2020).
2. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Curran Associates, Inc., 2017).
3. J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, 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)*, eds. J. Burstein, C. Doran and T. Solorio (Association for Computational Linguistics, Minneapolis, Minnesota, June 2019).
4. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, Large language models encode clinical knowledge, *Nature* **620**, 172 (2023).
5. K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis *et al.*, Toward expert-level medical question answering with large language models, *Nature medicine* **31**, 943 (2025).
6. H. Nori, N. King, S. M. McKinney, D. Carignan and E. Horvitz, Capabilities of gpt-4 on medical challenge problems, *arXiv preprint arXiv:2303.13375* **abs/2303.13375** (2023).
7. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* **36**, 1234 (02 2020).
8. R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon and T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, *Briefings in Bioinformatics* **23**, p. bbac409 (11 2022).
9. X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. G. Flores, Y. Zhang *et al.*, GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records, *arXiv preprint arXiv:2203.03540* **abs/2203.03540** (2022).
10. Z. Zhan and R. Zhang, Towards better multi-task learning: A framework for optimizing dataset combinations in large language models, in *Findings of the Association for Computational Linguistics: NAACL 2025*, eds. L. Chiruzzo, A. Ritter and L. Wang (Association for Computational Linguistics, Albuquerque, New Mexico, April 2025).
11. Z. Zhan, S. Zhou, H. Zhou, Z. Liu and R. Zhang, EPEE: Towards efficient and effective foundation models in biomedicine, *npj Health Systems* **3**, p. 30 (2026).
12. Z. Zhan, S. Zhou, M. Zeng, K. Yu, M. Song, X. Chen, J. Wang, Y. Hou and R. Zhang, Quantized large language models in biomedical natural language processing: Evaluation and recommendation, *arXiv preprint arXiv:2509.04534* **abs/2509.04534** (2025).
13. Z. Zhan, S. Zhou, X. Zhou, Y. Xiao, J. Wang, J. Deng, H. Zhu, Y. Hou, Y. Song, M. Lin and R. Zhang, Retrieval-augmented in-context learning for multimodal large language models in disease classification, *Journal of Biomedical Informatics* **178**, p. 105017 (2026).
14. Z. Zhan, M. Cui and R. Zhang, Can large language models self-correct in medical question an-

- swering? an exploratory study, *arXiv preprint arXiv:2604.00261* **abs/2604.00261** (2026).
15. K. Guu, K. Lee, Z. Tung, P. Pasupat and M. Chang, Retrieval augmented language model pre-training, in *Proceedings of the 37th International Conference on Machine Learning*, eds. H. D. III and A. Singh, Proceedings of Machine Learning Research, Vol. 119 (PMLR, 13–18 Jul 2020).
 16. V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen and W.-t. Yih, Dense passage retrieval for open-domain question answering, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, eds. B. Webber, T. Cohn, Y. He and Y. Liu (Association for Computational Linguistics, Online, November 2020).
 17. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel and D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in *Advances in Neural Information Processing Systems*, eds. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin (Curran Associates, Inc., 2020).
 18. G. Izacard and E. Grave, Leveraging passage retrieval with generative models for open domain question answering, in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, eds. P. Merlo, J. Tiedemann and R. Tsarfaty (Association for Computational Linguistics, Online, April 2021).
 19. S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. De Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. Rae, E. Elsen and L. Sifre, Improving language models by retrieving from trillions of tokens, in *Proceedings of the 39th International Conference on Machine Learning*, eds. K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu and S. Sabato, Proceedings of Machine Learning Research, Vol. 162 (PMLR, 17–23 Jul 2022).
 20. G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel and E. Grave, Atlas: Few-shot learning with retrieval augmented language models, *Journal of Machine Learning Research* **24**, 1 (2023).
 21. G. Xiong, Q. Jin, Z. Lu and A. Zhang, Benchmarking retrieval-augmented generation for medicine, in *Findings of the Association for Computational Linguistics: ACL 2024*, eds. L.-W. Ku, A. Martins and V. Srikumar (Association for Computational Linguistics, Bangkok, Thailand, August 2024).
 22. Z. Zhan, J. Wang, S. Zhou, J. Deng and R. Zhang, MMRAG: Multi-mode retrieval-augmented generation with large language models for biomedical in-context learning, *Journal of the American Medical Informatics Association* **32**, 1505 (10 2025).
 23. J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, Proximal policy optimization algorithms, *arXiv preprint arXiv:1707.06347* **abs/1707.06347** (2017).
 24. Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li *et al.*, Deepseek-math: Pushing the limits of mathematical reasoning in open language models, *arXiv preprint arXiv:2402.03300* **abs/2402.03300** (2024).
 25. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike and R. Lowe, Training language models to follow instructions with human feedback, in *Advances in Neural Information Processing Systems*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (Curran Associates, Inc., 2022).
 26. B. Jin, H. Zeng, Z. Yue, J. Yoon, S. O. Arik, D. Wang, H. Zamani and J. Han, Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning, *Second Conference on Language Modeling* **2025** (2025).
 27. R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders *et al.*, Webgpt: Browser-assisted question-answering with human feedback, *arXiv preprint arXiv:2112.09332* **abs/2112.09332** (2021).
 28. S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan and Y. Cao, React: Synergizing

- reasoning and acting in language models, *The Eleventh International Conference on Learning Representations* **2023** (2023).
29. T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda and T. Scialom, Toolformer: Language models can teach themselves to use tools, in *Advances in Neural Information Processing Systems*, eds. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine (Curran Associates, Inc., 2023).
 30. S. Zhou, Z. Xu, M. Zhang, C. Xu, Y. Guo, Z. Zhan, Y. Fang, S. Ding, J. Wang, K. Xu *et al.*, Large language models for disease diagnosis: A scoping review, *npj Artificial Intelligence* **1**, p. 9 (2025).
 31. Z. Zhan, Y. Hou, K. Yu, M. Zeng, A. Burgun, X. Chen and R. Zhang, An underexplored frontier: Large language models for rare disease patient education and communication—a scoping review, *arXiv preprint arXiv:2604.14179* **abs/2604.14179** (2026).
 32. D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang and P. Szolovits, What disease does this patient have? a large-scale open domain question answering dataset from medical exams, *Applied Sciences* **11** (2021).
 33. D. Vilares and C. Gómez-Rodríguez, HEAD-QA: A healthcare dataset for complex reasoning, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, eds. A. Korhonen, D. Traum and L. Màrquez (Association for Computational Linguistics, Florence, Italy, July 2019).
 34. Q. Jin, B. Dhingra, Z. Liu, W. Cohen and X. Lu, 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)*, eds. K. Inui, J. Jiang, V. Ng and X. Wan (Association for Computational Linguistics, Hong Kong, China, November 2019).
 35. S. Bedi, H. Cui, M. Fuentes *et al.*, Holistic evaluation of large language models for medical tasks with MedHELM, *Nature Medicine* **32**, 943 (2026).
 36. L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder and F. Wei, Text embeddings by weakly-supervised contrastive pre-training, *arXiv preprint arXiv:2212.03533* **abs/2212.03533** (2022).
 37. J. Johnson, M. Douze and H. Jégou, Billion-scale similarity search with gpus, *IEEE Transactions on Big Data* **7**, 535 (2021).
 38. E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, LoRA: Low-rank adaptation of large language models, *International Conference on Learning Representations* **2022** (2022).
 39. W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang and I. Stoica, Efficient memory management for large language model serving with pagedattention, in *Proceedings of the 29th Symposium on Operating Systems Principles*, , SOSP '23 Vol. 29 (Association for Computing Machinery, New York, NY, USA, 2023).
 40. J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le and D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in *Advances in Neural Information Processing Systems*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (Curran Associates, Inc., 2022).
 41. X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery and D. Zhou, Self-consistency improves chain of thought reasoning in language models, *The Eleventh International Conference on Learning Representations* **2023** (2023).
 42. E. B. Wilson, Probable inference, the law of succession, and statistical inference, *Journal of the American Statistical Association* **22**, 209 (1927).