

ChronoClinQA: A Longitudinal Multi-Admission Benchmark for Clinical Reasoning over Electronic Health Records

Tong Chen^{1,2,*}, Zimu Wang^{1,2,*}, Dongzhe Fan³, Luoyao Chen³, Haoyi Lian¹,
Yifan Wang³, Zhengni Yang^{1,2}, Zhengyong Jiang¹, Procheta Sen², Wei Wang¹,
Jionglong Su^{1,‡}, Qiaoyu Tan^{3,‡}

¹Xi'an Jiaotong-Liverpool University ²University of Liverpool

³New York University Shanghai

{Tong.Chen,Zimu.Wang}@liverpool.ac.uk, Qiaoyu.Tan@nyu.edu

Abstract

Large language models (LLMs) have shown strong performance on medical question answering benchmarks, yet existing evaluations remain limited in assessing longitudinal clinical reasoning across multiple hospital admissions. We introduce CHRONOCLINQA, a large-scale benchmark for longitudinal clinical reasoning grounded in multi-admission electronic health records. Built from longitudinal discharge summaries in MIMIC-IV-Note, CHRONOCLINQA contains 15,067 QA pairs spanning six clinically grounded reasoning dimensions and three complementary question formats: open-ended, adversarial, and unanswerable multiple-choice questions. Each instance requires reasoning over temporally accumulated patient trajectories, with contexts averaging over 16K tokens across up to eight hospital admissions. Experiments on both general-purpose and medical LLMs reveal substantial performance degradation in longitudinal and adversarial settings, highlighting persistent challenges in integrating cross-admission evidence and in temporal clinical reasoning. Our findings position CHRONOCLINQA as a challenging benchmark for advancing clinically grounded long-context reasoning in medical LLMs.

1 Introduction

Large language models (LLMs) have shown significant potential in medical applications, ranging from healthcare delivery (Mukherjee et al., 2024; Tu et al., 2024; Na et al., 2026a,b; Xu et al., 2026) to clinical decision support (Kim et al., 2024; Chai et al., 2025; Zhao et al., 2025; Fong et al., 2026). Despite achieving expert-level performance on licensing benchmarks (Chen et al., 2025; Zhou et al., 2025; Rao et al., 2026), current models remain limited in real-world clinical settings, as they struggle

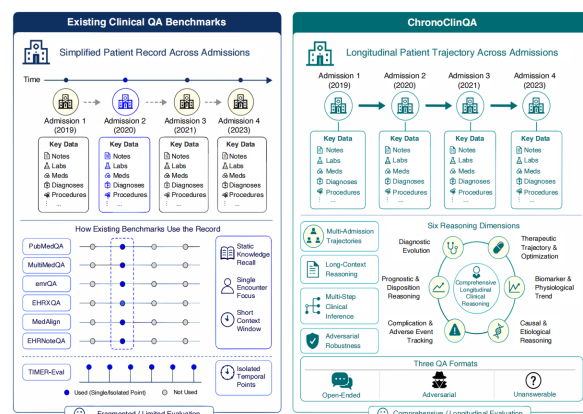


Figure 1: Comparison between existing clinical QA benchmarks and CHRONOCLINQA.

with *longitudinal, multi-step reasoning* across multiple hospital admissions.

Existing datasets on medical question answering (QA) are insufficient to evaluate these capabilities. As shown in Figure 1, traditional benchmarks, such as MedQA (Jin et al., 2020) and PubMedQA (Jin et al., 2019), mainly assess static recall of isolated textbook knowledge, lacking support for cross-admission reasoning. Electronic health record (EHR)-specific datasets, such as EHRXQA (Bae et al., 2023) and EHRNoteQA (Kweon et al., 2024) adopt extended task formats that either focus on single-visit records or have short context windows, offering little support for cross-admission integration or multi-step reasoning. TIMER-Eval (Cui et al., 2025) introduces time-aware QA over longitudinal EHRs but focuses on isolated temporal points rather than full longitudinal trajectories, leaving a gap in evaluating models’ ability to reason across multiple hospitalizations.

We present CHRONOCLINQA, a large-scale clinical reasoning dataset to address the aforementioned limitations. CHRONOCLINQA is constructed from longitudinal multi-admission discharge records in MIMIC-IV-Note (Johnson et al., 2023), and is built via a three-stage pipeline:

*Equal contribution.

†Work was done during the visit at NYU Shanghai.

‡Corresponding authors.

extracting structured clinical trajectories across consecutive hospitalizations, generating reasoning questions grounded in these trajectories, and filtering them through a format-adaptive quality adjudication framework. At its core is a six-dimensional clinical reasoning taxonomy: *Diagnostic Evolution*, *Therapeutic Trajectory and Optimization*, *Biomarker and Physiological Trend Reasoning*, *Causal and Etiological Reasoning*, *Complication and Adverse Event Tracking*, and *Prognostic and Disposition Reasoning* (see Figure 1), derived from established clinical decision-making frameworks (Rajpurkar et al., 2022; Newman-Toker et al., 2024) and operationalized via verifiable cross-admission changes. This design explicitly targets the types of longitudinal and multi-step reasoning that are absent in prior datasets, ensuring that each question challenges models to integrate evidence (Sun et al., 2026a,b) across temporal and clinical dimensions.

Our contributions are three-fold: **(1)** We introduce CHRONOCLINQA, a longitudinal clinical QA benchmark with 15,067 questions grounded in 7,000 multi-admission patient trajectories, targeting long-context temporal reasoning across six clinical dimensions. **(2)** We develop a scalable pipeline for constructing evidence-grounded longitudinal QA from discharge records through structured trajectory extraction, reasoning-oriented question generation, and quality adjudication. **(3)** We benchmark state-of-the-art general and medical LLMs, revealing substantial weaknesses in cross-admission evidence integration, adversarial reasoning, and long-context clinical inference.

2 Related Work

Clinical NLP Benchmarks have progressively evolved from standardized examination datasets (Jin et al., 2019, 2020) toward more clinically grounded resources (Singhal et al., 2022; Fleming et al., 2024), with recent frontier models achieving expert-level performance on licensing-style evaluations (Singhal et al., 2023; Nori et al., 2023). However, this progression has largely conflated clinical competence with performance on isolated, single-encounter snapshots of medical knowledge. Comprehensive surveys of LLMs in clinical settings (Thirunavukarasu et al., 2023) reveal that existing benchmarks systematically underrepresent the reasoning demands of real-world practice, where diagnostic and therapeutic decisions emerge from synthesizing fragmented evidence accumulated across

a patient’s longitudinal history. Efforts to base evaluation on real clinical notes (Fleming et al., 2024; Kweon et al., 2024) represent meaningful progress yet remain constrained by single-visit scope and limited scale, leaving the knowledge–practice gap fundamentally unaddressed.

Longitudinal EHR Modeling has been extensively studied in predictive tasks: readmission, mortality, and disease progression (Harutyunyan et al., 2019), where multi-visit information is encoded as temporal signals for supervised prediction rather than as evidence for interpretable reasoning. While emrQA (Pampari et al., 2018), EHRXQA (Bae et al., 2023), EHRNoteQA (Kweon et al., 2024) have made important strides toward free-text clinical reasoning, they have yet to model the continuous evolution of clinical trajectories across admissions. TIMER-Eval (Cui et al., 2025) begins to address temporal structure in EHRs but only targets isolated temporal points rather than the cross-admission dynamics that characterize complex longitudinal cases. Concurrently, benchmark design research has established that robustness evaluation requires both adversarial formulations (Rajpurkar et al., 2018) and unanswerable subsets to probe model calibration (Kadavath et al., 2022), dimensions absent from all existing resources.

3 Dataset Construction

We construct CHRONOCLINQA through a three-stage pipeline that transforms raw multi-admission discharges into a rigorously validated clinical reasoning benchmark. Starting from the MIMIC-IV-Note database (Johnson et al., 2023), we randomly sample 1,000 patient instances for each multi-visit setting spanning 2–8 admissions, yielding a diverse longitudinal cohort covering varying levels of temporal complexity. The pipeline combines structured longitudinal clinical representation, evidence-driven question seeding, and format-adaptive quality adjudication, with human supervision during construction. An overview is provided in Figure 2.

3.1 Longitudinal Clinical Representation

We first transform raw multi-admission discharge summaries into a unified structured representation that captures each patient’s clinical trajectory across hospitalizations. This stage is separated into three steps: rule-based structural decomposition, clinical note information extraction, and cross-admission clinical change detection.

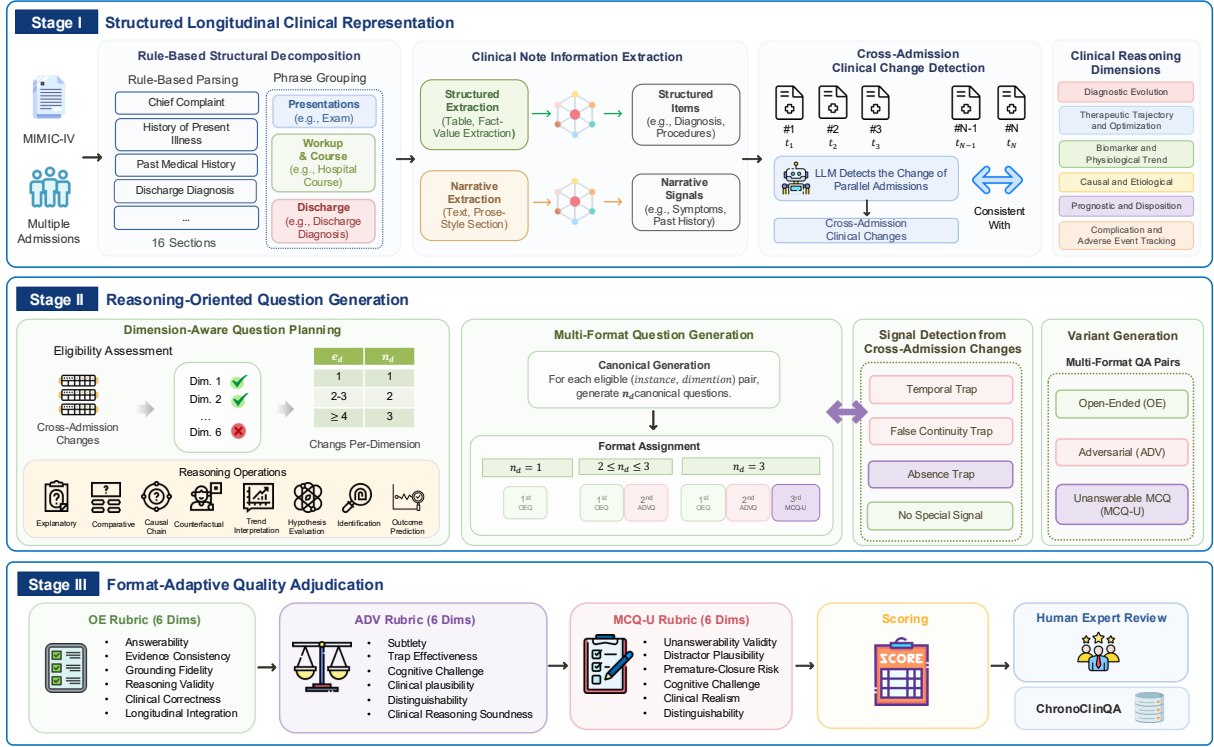


Figure 2: Overview of the CHRONOCLINQA construction pipeline.

Rule-Based Structural Decomposition MIMIC-IV discharge summaries follow a semi-structured format with identifiable section headers. We first apply rule-based segmentation to parse each note into 16 canonical clinical sections, which are further organized into three temporally ordered phases: *presentation*, *workup and course*, and *discharge*. This phase-level representation reflects the natural progression of a clinical encounter (Bowen, 2006) and enables scope-constrained downstream processing: each LLM receives only the phases relevant to its extraction objective, reducing irrelevant contextual noise. For patients with multiple admissions, notes are ordered chronologically and augmented with an explicit temporal timeline containing admission order, admission dates, length of stay, and inter-admission intervals, which is propagated throughout subsequent stages as a structured temporal context.

Clinical Note Information Extraction Some decomposed sections from the last step remain raw text. To enable efficient cross-admission comparison in Step 3, we extract structured clinical signals via DeepSeek-V4-Flash¹ per note, each targeting a distinct type of information. Structured extraction processes short, fact-dense sections (e.g., chief complaint, procedures, discharge diagnosis) to pro-

duce discrete clinical items such as diagnosis lists with primary/secondary labels. Narrative extraction processes longer, prose-style sections (e.g., history of present illness, hospital course, medication lists) to identify clinically meaningful signals embedded in narrative text, including clinical problems, medication changes with type and rationale, patient behaviors, and pending items. This ensures that structured clinical facts and narrative reasoning traces are captured without either interfering with the other’s extraction quality.

Cross-Admission Clinical Change Detection With the extracted information in place, we identify clinically meaningful changes across consecutive admission pairs. For each patient with $N \in [2, 8]$ admissions, we construct $N - 1$ consecutive pairs and submit each to DeepSeek-V4-Flash that compares compressed summaries of both notes, producing a structured set of cross-admission clinical changes. Each identified change is categorized according to six predefined clinically grounded reasoning dimensions:

- *Diagnostic Evolution*, capturing shifts in diagnosis status, certainty, or classification across admissions (Graber et al., 2005);
- *Therapeutic Trajectory and Optimization*, covering changes in treatments, medications, and clinical interventions (Spinewine et al., 2007);

¹hf.co/deepseek-ai/DeepSeek-V4-Flash

- *Biomarker and Physiological Trend Reasoning*, tracking directional changes in laboratory values and physiological parameters (Harutyunyan et al., 2019);
- *Causal and Etiological Reasoning*, linking events across admissions through explicit causal or etiological relationships (Hernan and Robins, 2020);
- *Complication and Adverse Event Tracking*, documenting newly emerged complications or adverse outcomes (Makary and Daniel, 2016);
- *Prognostic and Disposition Reasoning*, reflecting changes in functional status, discharge destination, or mortality trajectory (Krumholz, 2013).

These six clinically grounded dimensions span the inferential demands of longitudinal clinical reasoning; detailed definitions are provided in Appendix C.

3.2 Reasoning-Oriented Question Generation

Building on the cross-admission clinical changes identified in Stage 1, this stage systematically plans and generates a diverse set of reasoning questions. The process separates the decision of *what to ask* from *how to ask it*: a fully deterministic planning step first determines which clinical reasoning dimensions are supported and how many questions each warrants, after which an LLM-driven generation step produces questions in multiple formats.

Dimension-Aware Question Planning Not every patient record contains sufficient evidence to support all clinical dimensions. Therefore, before question generation, we assess the eligibility of each instance for every clinical dimension based on the cross-admission clinical changes identified in Section 3.1. For each eligible dimension d , the number of generated questions n_d is determined by its evidential richness e_d , defined as the number of clinically meaningful supporting changes. Specifically, $n_d \in 1, 2, 3$ corresponds to $e_d \in 1, [2-3], \geq 4$, respectively. The choice of requiring at least four evidence items, rather than three, for the maximum allocation of three questions is intentional: generating multiple questions from only three evidence signals can lead to over-extraction, producing clinically superficial questions that merely rephrase the same underlying observation from slightly different perspectives (Schuwirth and Van der Vleuten, 2011).

Beyond determining question quantity, the planning stage also specifies the reasoning modalities

to be elicited. Drawing on prior work in reasoning and question generation (Mihaylov et al., 2018; Tandon et al., 2018), together with principles from clinical assessment design (Daniel et al., 2019), we define eight reasoning operations (e.g., explanatory, comparative, counterfactual, etc.) that are paired with the corresponding clinical dimensions during question generation. Detailed definitions of these reasoning operations, along with their allocation strategies across the six clinical dimensions, are provided in Appendix D.

Multi-Format Question Generation The previous stage determines both the number of questions to generate for each clinical dimension and the reasoning operations to elicit, but not their presentation format. Following established principles in clinical assessment design (Epstein, 2007; Schuwirth and Van der Vleuten, 2011), we generate three complementary formats from a shared canonical representation: Open-Ended Questions (OEQ), Adversarial Questions (ADVQ), and Unanswerable Multiple-Choice Questions (MCQ-U).

Canonical Generation. For each eligible (instance, dimension) pair, DeepSeek-V4-Flash generates n_d canonical questions. Each canonical contains: (1) a longitudinal clinical scenario synthesized across admissions; (2) a core reasoning question requiring evidence from at least two time points; (3) a correct answer with an explicit reasoning chain; (4) supporting change indices for provenance tracing; and (5) clinically grounded common errors used for distractor construction. This canonical representation serves as a shared semantic anchor across all downstream formats, ensuring consistency in both reasoning requirements and gold-standard answers despite differences in surface form.

Format Assignment. Question formats are assigned through a deterministic, signal-conditioned strategy. Within each (instance, dimension) group of n_d canonicals, the first question is assigned OEQ (when $n_d = 1$). When $n_d \geq 2$, the second is assigned ADVQ; when $n_d = 3$, the third is assigned MCQ-U.

A *temporal trap* occurs when a clinical finding evolves from uncertainty to confirmation across admissions (Mamede et al., 2008). A *false continuity trap* arises when two changes share similar prior states but diverge into distinct outcomes, potentially masking different etiologies beneath su-

perificially similar findings (Norman, 2005). Either condition triggers assignment to ADVQ. By contrast, an *absence trap* occurs when the prior state of a clinical change is explicitly undocumented, creating a genuine information gap; such cases are assigned to MCQ-U. If no specialized signal is detected, the third question defaults to OEQ, since constructing clinically meaningful adversarial traps or distractors from marginal evidence often leads to implausible variants (Haladyna et al., 2002).

Variation Generation. Each canonical is subsequently transformed into its assigned format. OEQ variants are generated deterministically from canonical fields without additional LLM inference, with the reasoning chain directly serving as structured supervision (Wei et al., 2022). ADVQ variants embed the detected trap into an otherwise natural clinical question. Temporal traps encourage over-reliance on a single time-point observation (Croskerry, 2003; Mamede et al., 2008), whereas false-continuity traps frame superficially similar findings as manifestations of the same underlying condition (Norman, 2005; Kahneman, 2011). To preserve realism, no meta-linguistic cues indicating the presence of a trap are allowed. Finally, MCQ-U variants follow the unanswerable question paradigm of Rajpurkar et al. (2018): the correct answer explicitly states that the available notes are insufficient, while distractors correspond to plausible yet unsupported inferences that may arise from premature clinical closure (Krupat et al., 2017).

3.3 Format-Adaptive Quality Adjudication

To ensure the clinical validity and pedagogical reliability of the generated QA pairs, we adopt a format-adaptive LLM-as-a-judge framework using GPT-OSS-120B (OpenAI et al., 2025). Since different question formats impose distinct quality requirements, we design three specialized evaluation rubrics, each consisting of six dimensions scored on a 0–3 integer scale (total score: 0–18). For OEQ, the rubric evaluates answerability, evidence consistency, grounding fidelity, reasoning validity, clinical correctness, and longitudinal integration (Guyatt et al., 2008; Cook and Hatala, 2016). For ADVQ, the rubric focuses on trap effectiveness, subtlety, clinical reasoning soundness, cognitive challenge, clinical plausibility, and distinguishability (Croskerry, 2003; Kahneman, 2011). For MCQ-U, the evaluation assesses unanswerability validity, distractor plausibility, premature-closure

risk, cognitive challenge, clinical realism, and distinguishability (Rajpurkar et al., 2018; Jin et al., 2020). To validate adjudication reliability, we randomly sampled 100 instances per predefined reasoning dimension for manual comparison against model judgments, with a medical PhD student with clinical training experience and a clinically trained annotator averaging, yielding a Cohen’s Kappa (Cohen, 1960) of 73.19%.

A sample is accepted only if its total score is at least 15/18. In addition, any sample receiving a grounding fidelity score of ≤ 1 is automatically rejected regardless of its overall score, ensuring strict evidence grounding. Finally, we obtain a curated benchmark containing 15,067 high-quality clinical QA instances.

3.4 Dataset Characteristics

Overall, CHRONOCLINQA benchmark contains 7,000 subjects, 35,000 clinical notes, and 15,067 questions constructed over temporally accumulated patient histories. Unlike conventional clinical QA datasets that rely on isolated encounters, CHRONOCLINQA explicitly emphasizes *longitudinal temporal reasoning*: each subject contains 2–8 temporally ordered clinical notes (average 5.0). To evaluate model robustness under different reasoning settings, the benchmark further includes three complementary question formats: open-ended questions (67.5%), adversarial questions (16.9%), and uncertainty multiple-choice questions (MCQ-U, 15.6%). Detailed statistics are presented in Appendix B.

4 Experiments

4.1 Experimental Setup

We evaluate a diverse set of both *general-purpose* and *medical-domain* LLMs. Since many medical LLMs are developed via domain adaptation of existing foundation models, we organize the evaluated systems by their architectural lineage and clinical specialization. Specifically, we evaluate a set of general-purpose and medical LLMs spanning standard and reasoning-oriented architectures. General-purpose models include Qwen2.5-7B (Qwen et al., 2025), Qwen2.5-14B (Qwen et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), and Apertus-8B (Apertus et al., 2025), together with reasoning models Qwen3-8B (Yang et al., 2025), Hy3-295B², and Qwen3.5-397B-A17B (Qwen Team, 2026). We also include medical LLMs sharing similar

²huggingface.co/tencent/Hy3

Model	Reasoning	OEQ							ADVQ							MCQ-U	
		EmbedSim	R-L	BLEU	Clin-F1	Corr.	Comp.	RQ	EmbedSim	R-L	BLEU	Clin-F1	Corr.	Comp.	RQ	TA	Acc ↑
General-Purpose LLMs																	
Qwen2.5-7B	✗	82.96	22.31	5.36	21.96	91.47	72.13	88.87	77.93	16.95	3.42	17.42	50.23	30.58	48.16	49.70	71.07
Qwen3-8B	✓	80.62	22.41	6.28	22.52	95.60	86.75	94.75	74.85	16.67	3.82	17.14	57.40	45.73	56.80	56.93	43.40
Apertus-8B	✗	76.51	17.82	3.81	18.87	85.65	68.35	84.60	73.14	14.66	2.96	15.64	49.53	42.87	49.33	54.33	57.27
Llama-3.1-8B	✗	81.15	21.49	4.65	21.32	90.24	56.14	84.61	73.88	14.70	2.36	14.98	44.90	21.61	41.23	44.43	65.80
Qwen2.5-14B	✗	85.94	24.72	6.46	24.63	95.74	80.18	93.68	80.66	18.92	4.10	19.51	60.11	43.50	59.64	59.71	63.02
Hy3-295B	✓	85.01	24.09	6.23	23.13	92.85	91.80	92.85	74.22	15.79	3.09	16.29	62.60	61.27	62.67	79.27	37.35
Qwen3.5-397B-A17B	✓	87.55	25.10	6.26	24.61	97.75	96.15	97.80	82.42	19.86	4.87	20.82	76.00	74.93	76.27	83.33	63.33
Medical Domain LLMs																	
Meditron3-Qwen2.5-7B	✗	82.71	21.39	5.02	21.85	88.96	63.42	85.05	76.32	16.51	3.33	16.93	46.20	24.53	42.00	45.53	78.20
HuatuoGPT-o1-7B	✓	83.50	18.77	4.15	20.64	94.30	92.75	94.35	77.93	16.41	3.77	17.40	56.07	53.87	56.00	65.60	77.67
HuatuoGPT-o1-8B	✓	83.89	22.22	5.19	21.69	95.25	89.25	95.15	83.01	19.17	4.00	19.73	54.73	51.60	54.80	56.60	58.80
Meditron3-8B	✗	80.66	20.38	5.04	21.74	85.65	66.50	84.00	72.17	15.27	3.08	15.29	34.20	21.13	33.07	34.27	49.40
HuatuoGPT-3-8B	✓	84.91	20.41	4.43	19.69	92.80	91.85	92.80	76.03	14.95	3.12	14.38	53.47	52.93	53.47	66.13	64.53
Apertus-8B-MeditronFO	✗	67.92	14.11	2.49	13.61	73.25	53.10	72.05	63.53	9.78	1.35	9.37	36.53	27.60	36.00	49.60	69.93
Meditron3-Qwen2.5-14B	✗	83.15	22.64	5.50	22.69	94.55	81.75	94.30	74.41	15.37	2.89	15.77	52.33	41.80	52.13	60.00	56.00

Table 1: Experimental results on CHRONOCLINQA. OEQ is evaluated using embedding similarity (EmbedSim), ROUGE-L (R-L), BLEU, Clinical Entity F1 (Clin-F1), Correctness (Corr.), Completeness (Comp.), and Reasoning Quality (RQ). ADVQ is evaluated using EmbedSim, R-L, BLEU, Clin-F1, Correctness, Completeness, Reasoning Quality, and Trap Avoidance (TA). MCQ-U is evaluated using Accuracy (Acc). Higher values indicate better performance for all metrics. Best performance for each metric is highlighted in **bold**.

base architectures, including Meditron3-Qwen2.5-7B³, Meditron3-Qwen2.5-14B⁴, Meditron3-8B⁵, Apertus-8B-MeditronFO (Theimer-Lienhard et al., 2026), and the reasoning-oriented medical models HuatuoGPT-o1-7B, HuatuoGPT-o1-8B (Chen et al., 2024), and HuatuoGPT-3-8B⁶, which are designed to enhance multi-step clinical reasoning through explicit deliberation mechanisms. For fair comparison, the temperature is fixed at 0.0 for all models. Due to resource limitations, we randomly sample 5,000 data points in evaluation. Prompts used for each question format are provided in Appendix F.

Evaluation Metrics. For OEQ and ADVQ, we report **EmbedSim**, **ROUGE-L**, **BLEU**, and **Clinical Entity F1 (Clin-F1)**, together with LLM-as-a-judge metrics scored by GPT-4o-mini⁷. Specifically, **Correctness (Corr.)** and **Completeness (Comp.)** follow TIMER-Eval (Cui et al., 2025), while **Reasoning Quality (RQ)** is used for both formats and **Trap Avoidance (TA)** is additionally used for ADVQ. MCQ-U is evaluated using **Accuracy (Acc)**. These metrics assess semantic, lexical, clinical, reasoning, and adversarial robustness.

4.2 Experimental Results

Table 1 summarizes performance on OEQ, ADVQ, and MCQ-U. Overall, performance remains uneven across formats. Qwen3.5-397B-A17B leads

most on OEQ and ADVQ, yet its MCQ-U accuracy (63.33) trails the best result (78.20). This contrast shows that strong generation and rubric scores do not ensure reliable recognition of unanswerable clinical questions.

Performance on Open-Ended Questions. On OEQ, Qwen3.5-397B-A17B leads EmbedSim (87.55), ROUGE-L (25.10), Correctness (97.75), Completeness (96.15), and Reasoning Quality (97.80). Qwen2.5-14B achieves the best BLEU (6.46) and Clin-F1 (24.63). This pattern links scale and explicit reasoning to stronger semantic and rubric-based scores, but not uniform gains in lexical or entity overlap. The modest ROUGE-L, BLEU, and Clin-F1 scores further highlight the difficulty of aligning generated answers with precise longitudinal clinical evidence.

Medical Domain Adaptation Does Not Consistently Improve Performance. Medical-domain adaptation does not consistently outperform general-purpose models. Meditron3-Qwen2.5-7B trails Qwen2.5-7B on every OEQ and ADVQ metric, although it attains the highest MCQ-U accuracy (78.20). Meditron3-Qwen2.5-14B also underperforms its base across all metrics, whereas HuatuoGPT variants yield isolated gains. For example, HuatuoGPT-o1-8B records the highest ADVQ EmbedSim (83.01) without broad superiority. Domain specialization may therefore improve selected behaviors but does not reliably strengthen longitudinal evidence integration.

Performance on Adversarial Questions. At the aggregate level, all evaluated models obtain

³hf.co/OpenMeditron/Meditron3-Qwen2.5-7B
⁴hf.co/OpenMeditron/Meditron3-Qwen2.5-14B
⁵hf.co/EPFLiGHT/Meditron3-8B
⁶hf.co/FreedomIntelligence/HuatuoGPT-3-8B
⁷<https://developers.openai.com/api/docs/models/gpt-4o-mini>

lower semantic, lexical, entity-overlap, and rubric-based scores on ADVQ than on OEQ. For example, Qwen3.5-397B-A17B declines from 87.55 to 82.42 in EmbedSim, from 24.61 to 20.82 in Clin-F1, and from 97.80 to 76.27 in Reasoning Quality. Although it still leads seven of the eight ADVQ metrics, HuatuoGPT-o1-8B achieves the highest ADVQ EmbedSim (83.01). Because OEQ and ADVQ in Table 1 are disjoint, signal-conditioned question sets, these aggregate differences reflect both question-set composition and adversarial framing. Section 4.3 provides a matched analysis that reduces this content-selection confounding.

Performance on Unanswerable MCQ. MCQ-U accuracy ranges from 37.35 for Hy3-295B to 78.20 for Meditron3-Qwen2.5-7B, with HuatuoGPT-o1-7B close behind at 77.67. Reasoning-enabled models do not consistently dominate: Qwen3.5-397B-A17B reaches 63.33, whereas Qwen3-8B and Hy3-295B obtain 43.40 and 37.35, respectively. This spread indicates substantial variation in recognizing insufficient evidence and shows that strong OEQ or ADVQ performance does not ensure reliable decisions on unanswerable cases, underscoring persistent uncertainty-calibration challenges.

4.3 Paired Analysis OEQ–ADVQ Comparison

Because the OEQ and ADVQ results in Table 1 are computed on disjoint, signal-conditioned question sets, their aggregate differences may conflate adversarial framing with scenario selection. Therefore, we conducted a paired analysis in the 5,000-instance evaluation subset by aligning the OEQ and ADVQ variants derived from the same longitudinal clinical scenarios. The matched pool comprised 354 scenario pairs, with 343–354 valid pairs per model across the seven models included in this analysis. Within the matched subset, ADVQ reduced EmbedSim by 3.89–9.44 points and Clin-F1 by 2.91–5.49 points. All EmbedSim confidence intervals excluded zero, and the paired differences were significant under two-sided Wilcoxon signed-rank tests ($p \leq 0.015$ for EmbedSim and $p \leq 0.002$ for Clin-F1). For both metrics, all seven models showed significant OEQ–ADVQ differences after Holm correction across the 14 paired tests. The size of the paired effects was negligible to small (EmbedSim: $d_z = 0.18$ – 0.41 ; Clin-F1: $d_z = 0.19$ – 0.36), indicating a consistent but moderate degradation. HuatuoGPT-3-8B exhibited the largest EmbedSim reduction (9.44 points),

whereas Qwen2.5-14B achieved the highest absolute EmbedSim within the paired subset under both formats but still declined by 5.74 points. These findings show that the OEQ–ADVQ gap persists after matching longitudinal scenarios, providing more controlled evidence of sensitivity to adversarial formulation while avoiding a purely causal interpretation of the aggregate comparison.

4.4 Longitudinal Context Ablations

To distinguish longitudinal evidence integration from long-context effects, we evaluate Llama-3.1-8B and Meditron3-8B under three controlled context ablations. Latest-note-only and single-admission-only remove cross-admission evidence, whereas shuffled-order preserves the admission content and approximate context length but disrupts chronology. Across both models, their performances on OEQ and ADVQ are worse under all three ablations than under the full longitudinal setting. For Llama-3.1-8B, latest-note-only produces the largest reduction, decreasing EmbedSim from 81.15 to 74.66 on OEQ and from 73.88 to 65.14 on ADVQ; shuffled-order causes a smaller but consistent decline. Meditron3-8B is substantially more sensitive: shuffled-order reduces OEQ EmbedSim from 80.66 to 61.62, while latest-note-only reduces ADVQ EmbedSim from 72.17 to 53.52. Because shuffling preserves the available admissions while changing only their order, these results provide the clearest evidence that chronological organization contributes beyond context length or retrieval load alone. Nevertheless, the smaller shuffled-order effect for Llama-3.1-8B shows that sensitivity to chronology is model dependent.

MCQ-U exhibits a different pattern. Llama-3.1-8B improves from 65.80 under the full setting to 67.73–75.00 under the ablations, whereas Meditron3-8B declines from 49.40 to 4.28–23.07. Thus, reduced context does not uniformly impair unanswerable-question accuracy: it may remove distracting evidence for some models while causing others to make unsupported predictions under incomplete context. Overall, the generation results consistently demonstrate the value of cross-admission evidence, whereas the divergent MCQ-U results reveal substantial model-specific differences in uncertainty calibration. Additional analysis across context-length and temporal-depth bins is provided in Appendix A.

Model	Metric	OEQ	ADVQ	Δ_{pair}	95% CI	d_z	p
HuatuogPT-3-8B	EmbedSim	75.48	66.05	9.44	[7.03, 11.85]	0.41	< 0.001
	Clin-F1	19.46	13.97	5.49	[3.88, 7.10]	0.36	< 0.001
HuatuogPT-o1-7B	EmbedSim	72.22	67.41	4.81	[2.65, 6.97]	0.23	0.013
	Clin-F1	20.48	17.57	2.91	[1.31, 4.52]	0.19	0.002
Llama-3.1-8B	EmbedSim	72.74	65.74	7.00	[4.73, 9.27]	0.32	< 0.001
	Clin-F1	20.63	15.48	5.15	[3.27, 7.02]	0.29	< 0.001
Meditron3-Qwen2.5-7B	EmbedSim	73.51	67.98	5.54	[3.25, 7.83]	0.25	0.004
	Clin-F1	21.56	16.93	4.63	[2.86, 6.39]	0.28	< 0.001
Qwen2.5-14B	EmbedSim	77.28	71.54	5.74	[3.75, 7.72]	0.30	< 0.001
	Clin-F1	24.47	19.58	4.89	[3.13, 6.66]	0.29	< 0.001
Qwen2.5-7B	EmbedSim	73.48	69.59	3.89	[1.62, 6.15]	0.18	0.015
	Clin-F1	21.99	18.11	3.88	[2.10, 5.67]	0.23	< 0.001
Qwen3-8B	EmbedSim	71.04	64.15	6.89	[4.10, 9.68]	0.26	< 0.001
	Clin-F1	21.70	16.58	5.12	[3.11, 7.13]	0.27	< 0.001

Table 2: Paired OEQ–ADVQ comparison on matched longitudinal clinical scenarios from the 5,000-instance evaluation subset. EmbedSim, Clin-F1, and their paired differences are reported on a 0–100 scale. $\Delta_{\text{pair}} = \text{OEQ} - \text{ADVQ}$, such that a positive value indicates degradation under ADVQ. Confidence intervals are 95% CIs for the mean paired difference, d_z is the paired standardized mean difference, and p is obtained from a two-sided Wilcoxon signed-rank test. All comparisons remain significant after Holm correction across the 14 paired tests.

Model	Setting	OEQ				ADVQ				MCQ-U
		EmbedSim	R-L	BLEU	Clin-F1	EmbedSim	R-L	BLEU	Clin-F1	Acc
Llama-3.1-8B	Full	81.15	21.49	4.65	21.32	73.88	14.70	2.36	14.98	65.80
	Latest	74.66	18.46	3.64	17.37	65.14	11.09	1.49	10.23	75.00
	Single	77.98	20.11	4.17	19.10	69.34	12.90	1.93	12.72	70.07
	Shuffled	79.20	20.82	4.45	20.57	70.61	13.57	2.19	13.48	67.73
Meditron3-8B	Full	80.66	20.38	5.04	21.74	72.17	15.27	3.08	15.29	49.40
	Latest	66.89	13.96	3.27	15.27	53.52	8.00	1.93	9.05	4.28
	Single	68.90	14.31	3.28	17.12	61.08	10.73	2.55	12.73	14.87
	Shuffled	61.62	10.62	2.42	14.63	58.15	8.39	2.01	12.85	23.07

Table 3: Controlled longitudinal-context ablations for Llama-3.1-8B and Meditron3-8B. *Full* preserves all admissions in chronological order; *Latest* retains only the most recent admission; *Single* retains one admission; and *Shuffled* preserves the same admission content while permuting admission order. Scores are computed over valid outputs and reported on a 0–100 scale. The best result within each model and metric is highlighted in **bold**.

4.5 Error Analysis

To better understand model limitations on CHRON-OCLINQA, we qualitatively analyzed representative prediction failures from both general-purpose and medical-domain LLMs. As summarized in Table 8, we identify four recurring failure modes that consistently impair performance across OEQ, ADVQ, and MCQ-U settings:

Temporal Shortcut Reasoning. Models often rely on temporally local or salient evidence while underutilizing longitudinal context. As shown in Table 8, when predicting a disposition transition from home service to an extended care facility, the model attributed the change to chronic psychiatric or cognitive conditions rather than acute hospitalization-related functional decline (e.g., fracture and mobility impairment). This sug-

gests a tendency to prioritize prominent historical conditions over temporally distributed evidence.

Over-Reliance on Salient Clinical Entities.

Models also overfit to clinically salient entities such as medications, pathogens, or disease patterns. In a recurrent urinary tract infection case, the model attributed an earlier admission to *Proteus mirabilis* despite no supporting microbiology evidence, likely relying on stereotypical associations between catheterization and infection patterns rather than record-grounded evidence.

Failure to Abstain Under Clinical Uncertainty.

In MCQ-U settings, models frequently fail to recognize insufficient evidence and instead commit to plausible but unsupported conclusions. For example, when mild liver enzyme elevation occurred during concurrent antibiotic use, the model attributed

hepatotoxicity to a specific drug despite the absence of evidence for causal attribution. This behavior aligns with the low accuracy in Table 1, indicating poor uncertainty calibration under ambiguous evidence.

Distractor Susceptibility in Adversarial Settings.

Adversarial questions expose a tendency to follow clinically plausible but misleading cues. In one case, transient transaminitis was misattributed to antibiotic-induced liver injury, whereas longitudinal evidence more strongly supported infection-related metabolic instability. These failures suggest that strong performance on CHRONOCLINQA requires not only medical knowledge, but also temporal evidence integration, causal reasoning, distractor resistance, and calibrated uncertainty.

5 Conclusion

We introduce CHRONOCLINQA, a large-scale benchmark for evaluating longitudinal clinical reasoning across multiple hospital admissions. ChronoClinQA combines long-context patient trajectories, six clinically grounded reasoning dimensions, and three complementary QA formats to assess multi-step temporal reasoning and robustness in realistic clinical settings. Through a three-stage construction pipeline with format-adaptive quality adjudication, we curate 15,067 high-quality QA pairs spanning diverse longitudinal scenarios. Experimental results show that even strong general-purpose and medical LLMs struggle with cross-admission reasoning, adversarial clinical traps, and unanswerable settings, highlighting substantial gaps between current model capabilities and real-world longitudinal clinical reasoning demands.

Limitations

Despite its scale and clinical grounding, CHRONOCLINQA has two main limitations. **(1)** The benchmark is constructed from the single MIMIC-IV-Note (Johnson et al., 2023) dataset, which is, to the best of our knowledge, the only publicly accessible, de-identified clinical note corpora that provide large-scale discharge summaries linkable across patient admissions, making them particularly suitable for longitudinal multi-admission QA. **(2)** Due to resource limitations, we only evaluate two of the latest frontier-scale proprietary models. Nevertheless, the evaluated models cover a diverse set of general-purpose and medical LLMs, including both standard and reasoning-oriented architectures

with different model lineages and parameter scales, which we believe is sufficient to reveal consistent weaknesses in cross-admission evidence integration, adversarial robustness, long-context reasoning, and uncertainty calibration.

Ethical Considerations

We discuss the following ethical considerations related to our CHRONOCLINQA dataset as follows:

(1) Data Access and Privacy. CHRONOCLINQA is constructed from MIMIC-IV-Note, which is distributed under the PhysioNet Credentialed Health Data License 1.5.0 and governed by the PhysioNet Credentialed Health Data Use Agreement (DUA) 1.5.0. We adhere to the license and restrict CHRONOCLINQA to research use only. Access is restricted to users who have independently obtained authorized access to MIMIC-IV-Note, completed the required training, and signed the applicable PhysioNet Data Use Agreement. Since MIMIC-IV-Note is de-identified by the original data providers, all benchmark construction and evaluation are conducted exclusively on anonymized records, and no re-identification attempts are performed. **(2) Intended Use.** CHRONOCLINQA is intended for research on longitudinal clinical reasoning over electronic health records and for benchmarking medical LLMs. The benchmark may support the development of more reliable and evidence-grounded models for longitudinal EHR reasoning. Researchers may also adapt our dataset construction methodology to develop related clinical reasoning benchmarks. **(3) Potential Risks and Mitigation.** Although the source data are de-identified, CHRONOCLINQA may inherit biases from historical ICU-centered clinical records and institutional practices in MIMIC-IV-Note. We therefore conduct manual spot-checking on randomly sampled instances to assess annotation quality, privacy-sensitive content, and potential dataset issues. We do not anticipate substantial additional privacy risks beyond those associated with the original dataset. **(4) Use of AI.** We used ChatGPT for language refinement and figure polishing. All AI-assisted outputs were manually reviewed and verified by the authors, who retain full responsibility for the final content.

References

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina

- Messmer, Dhia Garbaya, Eduard Frank Durech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, and 84 others. 2025. [Apertus: Democratizing open and compliant llms for global language environments](#). *Preprint*, arXiv:2509.14233.
- Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and Edward Choi. 2023. [Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 3867–3880. Curran Associates, Inc.
- Judith L Bowen. 2006. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21):2217–2225.
- Jiajun Chai, Sicheng Li, Yuqian Fu, Dongbin Zhao, and Yuanheng Zhu. 2025. [Empowering llm agents with zero-shot optimal decision-making through q-learning](#). In *International Conference on Learning Representations*, volume 2025, pages 78319–78365.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. [Huatuogpt-o1, towards medical complex reasoning with llms](#). *Preprint*, arXiv:2412.18925.
- Tong Chen, Zimu Wang, Yiyi Miao, Haoran Luo, Yuanfei Sun, Wei Wang, Zhengyong Jiang, Procheta Sen, and Jionglong Su. 2025. [MedFact: A large-scale Chinese dataset for evidence-based medical fact-checking of LLM responses](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32340–32353, Suzhou, China. Association for Computational Linguistics.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- David A Cook and Rose Hatala. 2016. Validation of educational assessments: a primer for simulation and beyond. *Advances in simulation*, 1(1):31.
- Pat Croskerry. 2003. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic medicine*, 78(8):775–780.
- Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam Shah. 2025. [Timer: Temporal instruction modeling and evaluation for longitudinal clinical records](#). *Preprint*, arXiv:2503.04176.
- Michelle Daniel, Joseph Rencic, Steven J Durning, Eric Holmboe, Sally A Santen, Valerie Lang, Temple Ratcliffe, David Gordon, Brian Heist, Stuart Lubarsky, and 1 others. 2019. Clinical reasoning assessment methods: a scoping review and practical guidance. *Academic Medicine*, 94(6):902–912.
- Ronald M Epstein. 2007. Assessment in medical education. *New England journal of medicine*, 356(4):387–396.
- Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gavidis, Scott Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, and 11 others. 2024. [Medalign: A clinician-generated dataset for instruction following with electronic medical records](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22021–22030.
- Stephanie Fong, Zimu Wang, Guilherme C Oliveira, Xiangyu Zhao, Yiwen Jiang, Jiahe Liu, Beau-Luke Colton, Scott W. Woods, Martha Shenton, Barnaby Nelson, Zongyuan Ge, and Dominic Dwyer. 2026. [CHiRPE: A step towards real-world clinical NLP with clinician-oriented model explanations](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 646–658, Rabat, Morocco. Association for Computational Linguistics.
- Mark L Graber, Nancy Franklin, and Ruthanna Gordon. 2005. Diagnostic error in internal medicine. *Archives of internal medicine*, 165(13):1493–1499.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Gordon H Guyatt, Andrew D Oxman, Gunn E Vist, Regina Kunz, Yngve Falck-Ytter, Pablo Alonso-Coello, and Holger J Schünemann. 2008. Grade: an emerging consensus on rating quality of evidence and strength of recommendations. *Bmj*, 336(7650):924–926.
- Thomas M Haladyna, Steven M Downing, and Michael C Rodriguez. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333.
- Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96.
- MA Hernan and JM Robins. 2020. Causal inference: What if chapman hall/crc, boca raton.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. [What disease does this patient have? a large-scale open domain question answering dataset from medical exams](#). *Preprint*, arXiv:2009.13081.

- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. [MIMIC-IV-Note: Deidentified free-text clinical notes](#). *PhysioNet*. Version 2.2.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. [Mdagents: An adaptive collaboration of llms for medical decision-making](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452. Curran Associates, Inc.
- Harlan M Krumholz. 2013. Post-hospital syndrome—a condition of generalized risk. *The New England journal of medicine*, 368(2):100.
- Edward Krupat, Jolie Wormwood, Richard M Schwartzstein, and Jeremy B Richards. 2017. Avoiding premature closure and reaching diagnostic accuracy: some key predictive factors. *Medical education*, 51(11):1127–1137.
- Sunjun Kweon, Jiyou Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jee-won Yang, Seunghyun Won, and Edward Choi. 2024. [Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 124575–124611. Curran Associates, Inc.
- Martin A Makary and Michael Daniel. 2016. Medical error—the third leading cause of death in the us. *Bmj*, 353.
- Silvia Mamede, Henk G Schmidt, Remy MJP Rikers, Júlio César Penaforte, and João Macedo Coelho-Filho. 2008. Influence of perceived difficulty of cases on physicians’ diagnostic reasoning. *Academic Medicine*, 83(12):1210–1216.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2381–2391.
- Subhabrata Mukherjee, Paul Gamble, Markel Sanz Ausin, Neel Kant, Kriti Aggarwal, Neha Manjunath, Debajyoti Datta, Zhengliang Liu, Jiayuan Ding, Sophia Busacca, Cezanne Bianco, Swapnil Sharma, Rae Lasko, Michelle Voisard, Sanchay Harneja, Darya Filippova, Gerry Meixiong, Kevin Cha, Amir Youssefi, and 7 others. 2024. [Polaris: A safety-focused llm constellation architecture for healthcare](#). *Preprint*, arXiv:2403.13313.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. [Overview of the PsyDefDetect shared task at BioNLP 2026: Detecting levels of psychological defense mechanisms in supportive conversations](#). In *BioNLP 2026*, pages 932–943, San Diego, California. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. [You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 14428–14448, San Diego, California, United States. Association for Computational Linguistics.
- David E Newman-Toker, Najlla Nassery, Adam C Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D Clemens, Zheyu Wang, Yuxin Zhu, Ali S Saber Tehrani, Mehdi Fanai, Ahmed Hassoon, and 1 others. 2024. Burden of serious harms from diagnostic error in the usa. *BMJ Quality & Safety*, 33(2):109–120.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. [Capabilities of gpt-4 on medical challenge problems](#). *Preprint*, arXiv:2303.13375.
- Geoffrey Norman. 2005. Research in clinical reasoning: past history and current trends. *Medical education*, 39(4):418–427.
- OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. [emrQA: A large corpus for question answering on electronic medical records](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan

- Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Qwen Team. 2026. [Qwen3.5: Towards native multimodal agents](#).
- Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. Ai in health and medicine. *Nature medicine*, 28(1):31–38.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.
- Arya S. Rao, Kaiz P. Esmail, Richard S. Lee, Sharon Jiang, Bianca Arraiza Carlo, Jasleen Gill, Praneet Khanna, Ezra Kalmowitz, Basile Montagnese, Kimia Heydari, Qiao Jiao, Ethan Bott, Dan Nguyen, Grace Wang, Michael Hood, Adam B. Landman, and Marc D. Succi. 2026. [Large language model performance and clinical reasoning tasks](#). *JAMA Network Open*, 9(4):e264003–e264003.
- Lambert WT Schuwirth and Cees PM Van der Vleuten. 2011. Programmatic assessment: from assessment of learning to assessment for learning. *Medical teacher*, 33(6):478–485.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, and 11 others. 2022. [Large language models encode clinical knowledge](#). *Preprint*, arXiv:2212.13138.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaeckermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023. [Towards expert-level medical question answering with large language models](#). *Preprint*, arXiv:2305.09617.
- Anne Spinewine, Kenneth E Schmader, Nick Barber, Carmel Hughes, Kate L Lapane, Christian Swine, and Joseph T Hanlon. 2007. Appropriate prescribing in elderly people: how well can it be measured and optimised? *The Lancet*, 370(9582):173–184.
- Yuanfu Sun, Kang Li, Dongzhe Fan, Jiajin Liu, and Qiaoyu Tan. 2026a. Agentgl: Towards agentic graph learning with llms via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25313–25335.
- Yuanfu Sun, Kang Li, Pengkang Guo, Jiajin Liu, and Qiaoyu Tan. 2026b. Mario: Multimodal graph reasoning with large language models. *arXiv preprint arXiv:2603.05181*.
- Niket Tandon, Bhavana Dalvi, Joel Grus, Wen-tau Yih, Antoine Bosselut, and Peter Clark. 2018. Reasoning about actions and state changes by injecting commonsense knowledge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 57–66.
- Xavier Theimer-Lienhard, Mushtaha El-Amin, Fay Elhassan, Sahaj Vaidya, Victor Cartier-Negadi, David Sasu, Lars Klein, and Mary-Anne Hartley. 2026. [Fully open meditron: An auditable pipeline for clinical llms](#). *Preprint*, arXiv:2605.16215.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, and 14 others. 2024. [Towards generalist biomedical ai](#). *NEJM AI*, 1(3):AIoa2300138.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Qingyang Xu, Yaling Shen, Stephanie Fong, Zimu Wang, Yiwen Jiang, Xiangyu Zhao, Jiahe Liu, Zhongxing Xu, Vincent Lee, and Zongyuan Ge. 2026. [Do no harm: Exposing hidden vulnerabilities of llms via persona-based client simulation attack in psychological counseling](#). *Preprint*, arXiv:2604.04842.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Xiangyu Zhao, Yaling Shen, Yiwen Jiang, Zimu Wang, Jiahe Liu, Maxmartwell H Cheng, Guilherme C Oliveira, Robert Desimone, Dominic Dwyer, and Zongyuan Ge. 2025. [It hears, it sees too: Multimodal llm for depression detection by integrating visual understanding into audio language models](#). *Preprint*, arXiv:2511.19877.
- Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna M. Espinoza, Lindsay Welton, Xinyan Mai, Yanwei Jin, Zidu Xu, Yuen-Hei Chung, Yiyun Xing, Meng-Han Tsai, Emily B Schaffer, Yucheng Shi, Ninghao Liu, Zirui Liu, and

A Impact of Context Length and Temporal Depth

To better understand how longitudinal clinical complexity is associated with model performance, we analyze results across different temporal depths and context lengths, as shown in Figure 3. We evaluate selected general-purpose and medical-domain LLMs using ROUGE-L on OEQ and accuracy on MCQ-U.

Longer Contexts Do Not Consistently Improve Performance. Across the context-length bins, several models obtain lower scores at 24K–32K tokens than at 4K–8K tokens, particularly in MCQ-U accuracy. However, this pattern is neither monotonic nor universal. For example, Qwen2.5-7B exhibits a ROUGE-L rebound in the longest context bin, while HuatuoGPT-o1-7B and Meditron3-Qwen2.5-7B achieve their highest accuracy at intermediate context lengths. These results therefore suggest that additional context does not consistently translate into better performance and may be associated with degradation for some models and metrics. This pattern may reflect the greater demands that longer contexts place on identifying, filtering, and integrating clinically relevant information across encounters.

Greater Temporal Depth Produces Non-Monotonic Performance. Performance also varies non-monotonically as the number of clinical notes increases. No model improves consistently across all temporal depths. Several accuracy curves decline at intermediate note counts and subsequently recover, whereas ROUGE-L generally fluctuates within a narrower range. These patterns suggest that performance becomes less predictable as more encounters are introduced. They are consistent with the possibility that models must reconcile temporally distributed or evolving clinical information. However, because note count is only a proxy for temporal reasoning depth and may also correlate with context length and case composition, the observed changes cannot be attributed exclusively to temporal reasoning difficulty.

Medical Adaptation Provides Selective Rather Than Uniform Advantages. Medical-domain

models, particularly Meditron3-Qwen2.5-7B and HuatuoGPT-o1-7B, achieve relatively high accuracy under several long-context and high-note-count settings. Nevertheless, their performance remains non-monotonic, and their advantages do not consistently extend to ROUGE-L or to every context range. The figure therefore provides evidence of selective advantages under some longitudinal settings rather than uniformly greater robustness. Overall, domain specialization alone does not appear sufficient to ensure reliable performance as context length and temporal depth increase.

B Detailed Dataset Statistics

Statistic	Value
<i>Dataset Scale</i>	
# Subjects	7,000
# Questions	15,067
# Discharge Notes	35,000
Notes per Subject	5.0 ± 2.0
Questions per Subject	2.5 ± 1.3
<i>Context Length</i>	
Context Words ($\mu \pm \sigma$)	$9,040 \pm 3,422$
Context Tokens ($\mu \pm \sigma$)	$16,143 \pm 6,269$
<i>Long-Context Pressure (% Subjects)</i>	
> 4K tokens	100.0%
> 8K tokens	99.7%
> 16K tokens	43.0%
> 32K tokens	1.7%
<i>Question Format (# / %)</i>	
Open-ended	10,163 / 67.5%
Adversarial	2,551 / 16.9%
MCQ-U	2,353 / 15.6%

Table 4: Overview statistics of CHRONOCLINQA.

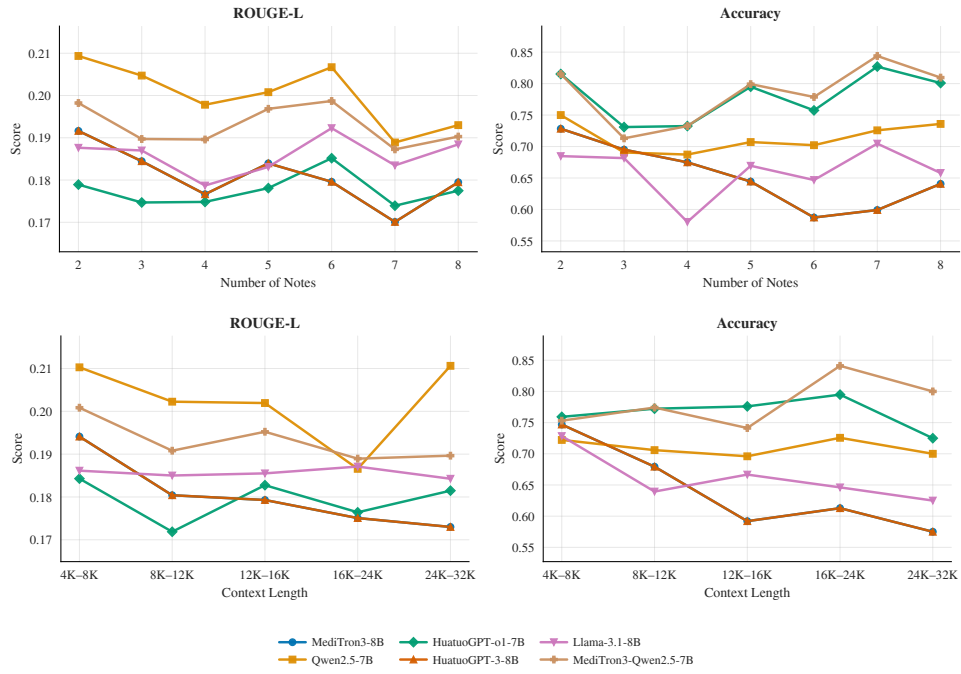


Figure 3: Model performance across temporal depth and context length bins. We report ROUGE-L for free-form generation and Accuracy for MCQ-U.

C Clinical Dimension Definition

Clinical Dimension	Definition	Longitudinal Signal	Example Reasoning Focus
Six-Dimensional Clinical Reasoning Taxonomy			
<i>Diagnostic Evolution</i>	Reasoning about changes in diagnoses across admissions, including emergence, resolution, refinement, or shifts in diagnostic certainty.	Changes in diagnosis status, severity, certainty, or classification over time.	How or why a diagnosis changed between admissions; identifying disease progression or diagnostic revision.
<i>Therapeutic Trajectory and Optimization</i>	Reasoning about treatment adjustments and clinical interventions across longitudinal care.	Medication changes, treatment escalation/de-escalation, intervention responses, or optimization decisions.	Why therapies were modified, discontinued, or intensified over time.
<i>Biomarker and Physiological Trend Reasoning</i>	Reasoning over temporal trends in laboratory findings and physiological measurements.	Directional changes in biomarkers, vital signs, or physiological indicators across visits.	Interpreting worsening, recovery, or clinically meaningful trends in patient status.
<i>Causal and Etiological Reasoning</i>	Reasoning about causal relationships linking events, conditions, or interventions across admissions.	Sequential clinical events with explicit or implied causal dependency.	Identifying causes of complications, disease progression, or treatment outcomes.
<i>Complication and Adverse Event Tracking</i>	Reasoning about newly emerged complications, adverse outcomes, or secondary clinical events.	Emergence or progression of complications and treatment-related adverse effects.	Recognizing downstream clinical consequences or adverse developments.
<i>Prognostic and Disposition Reasoning</i>	Reasoning about likely clinical outcomes, discharge planning, and functional prognosis.	Changes in discharge destination, functional status, readmission risk, or mortality trajectory.	Predicting future clinical course or explaining disposition decisions.

Table 5: Definitions of the six clinically grounded reasoning dimensions used in CHRONOCLINQA. Each dimension operationalizes a distinct type of longitudinal reasoning signal derived from cross-admission clinical changes and serves as a basis for question generation.

D Question Planning

Mode	Description
Explanatory	Why a clinical decision was made; articulating the reasoning process
Comparative	Contrasting clinical states or decisions across different time points
Counterfactual	What would have occurred under a different clinical decision
Trend Interpretation	Clinical significance of a longitudinal change in biomarkers or physiology
Causal Chain	Causal relationships between sequential clinical events
Etiological	Underlying cause or origin of a clinical condition
Identification	Recognizing or detecting a specific clinical finding or complication
Outcome Prediction	Likely clinical trajectory given the patient's course

Table 6: The eight reasoning operations defined for CHRONOCLINQA, ordered by cognitive complexity from pattern recognition to evaluative judgment.

Dimension	Base Mode ($e = 1$)	Expanded Modes ($e \geq 2$)	Reference
Diagnostic Evolution	Explanatory	Comparative	(Graber et al., 2005)
Therapeutic Trajectory and Optimization	Explanatory	Comparative; Counterfactual ($e \geq 3$)	(Spinewine et al., 2007)
Biomarker and Physiological Trend	Trend Interpretation	Comparative	(Harutyunyan et al., 2019)
Causal and Etiological	Causal Chain	Etiological ($e \geq 3$)	(Hernan and Robins, 2020)
Complication and Adverse Event Tracking	Identification	Causal Chain	(Makary and Daniel, 2016)
Prognostic and Disposition	Outcome Prediction	Explanatory	(Krumholz, 2013)

Table 7: Per-dimension reasoning operation allocation. The base operation is assigned when evidence is minimal ($e = 1$), while expanded modes are enabled as evidence accumulates ($e \geq 2$ or $e \geq 3$).

E Error Analysis Examples

Error Type	Clinical Scenario	Model Prediction	Evidence in Notes	Ground Truth	Observed Failure
Temporal Shortcut Reasoning	Disposition changed from home service to an extended care facility after a fall-related hospitalization.	<i>Disposition change due to dementia or psychiatric decline</i> (Llama-3.1-8B)	Patient developed a left humerus fracture and bilateral upper-extremity swelling with reduced mobility; rehabilitation and assisted care were recommended.	Functional decline caused by fracture and hematoma required rehabilitation support.	The model overemphasizes salient prior conditions (e.g., dementia) and ignores longitudinal evidence indicating that acute functional decline drove the disposition transition.
Over-Reliance on Salient Clinical Entities	Patient with recurrent UTIs and suprapubic catheterization later developed obstructive urosepsis. The task asked for the organism in an earlier admission.	<i>Proteus mirabilis</i> (HuatuoGPT-o1-7B, Qwen2.5-7B)	The clinical notes contain no microbiology result or urine culture evidence for the earlier hospitalization.	Cannot be determined from the clinical notes.	The model relies on medical heuristics (catheter + stones → <i>Proteus</i>) rather than evidence explicitly documented in the notes.
Failure to Abstain Under Uncertainty	Mild ALT elevation occurred while multiple antibiotics were simultaneously administered.	<i>Zosyn caused the ALT elevation</i> (Llama-3.1-8B)	The notes do not provide timing, comparative hepatotoxicity evidence, or diagnostic support to attribute causality to one specific antibiotic.	Cannot be determined from available clinical information.	Instead of abstaining when evidence is insufficient, the model commits to a specific answer, reflecting poor uncertainty calibration.
Distractor Susceptibility	Transient transaminitis occurred during infection-related admissions with sepsis-like physiology.	<i>Antibiotic-induced liver injury</i> (Llama-3.1-8B)	Liver enzymes were normal in prior and later admissions, while transient elevation appeared only during acute metabolic instability.	Transient ischemic hepatitis (shock liver) secondary to hypoperfusion/sepsis.	The model follows clinically plausible distractors (infection + antibiotics) while ignoring longitudinal temporal evidence supporting reversible ischemic injury.

Table 8: Representative error cases on CHRONOCLINQA: temporal shortcut reasoning, over-reliance on salient clinical entities, failure to abstain under uncertainty, and distractor susceptibility.

F Prompt

Structured Clinical Extraction Prompt

You are a clinical text extraction specialist. Your task is to extract factual items from short clinical sections.

=== Chief Complaint ===

{chief_complaint}

=== Procedures Performed ===

{procedures}

=== Discharge Diagnosis ===

{discharge_diagnosis}

=== Admission Physical Exam ===

{admission_exam}

=== Discharge Physical Exam ===

{discharge_exam}

Rules:

- Copy information verbatim from the source text.
- Extract discharge diagnoses as separate items and preserve PRIMARY/SECONDARY labels if explicitly provided.
- Split examination findings according to body-system headers (e.g., HEENT, CARDIAC).
- Store non-diagnostic statements (e.g., expired, passed away) separately.

Respond strictly in JSON format:

```
{
  chief_complaint, procedures, discharge_diagnosis,
  admission_exam_findings, discharge_exam_findings
}
```

Figure 4: Representative prompt for structured extraction from short, fact-dense clinical sections.

Narrative Clinical Extraction Prompt

You are a clinical text extraction specialist. Your task is to extract structured signals from a patient's clinical narrative.

=== History of Present Illness ===

{hpi}

=== Brief Hospital Course ===

{hospital_course}

=== Medications on Admission ===

{meds_admission_raw}

=== Discharge Medications ===

{meds_discharge_raw}

Rules:

- Extract explicitly mentioned clinical problems only.
- Identify medication changes by comparing admission and discharge medications (e.g., ADDED, STOPPED, DOSE_INCREASED).
- Extract patient actions, symptoms, temporal expressions, and pending items only when explicitly supported by source text.
- Preserve supporting rationale and evidence as verbatim quotations.

Respond strictly in JSON format:

```
{
  problems, medication_changes, patient_active_actions,
  symptoms_reported, time_expressions, pending_at_discharge
}
```

Figure 5: Representative prompt for narrative clinical extraction from narrative sections.

Cross-Admission Clinical Change Detection Prompt

You are a clinical longitudinal reasoning specialist. Compare two admissions for the SAME patient and identify factual changes (deltas) between them.

Admission 1 ({n1_admission} to {n1_discharge}):

{n1_summary}

Admission 2 ({n2_admission} to {n2_discharge}),
{gap_days} days after admission 1 discharge:
{n2_summary}

Six Delta Types (Aligned with Clinical Reasoning Dimensions)

1. D1 Diagnostic Evolution

A diagnosis is established, refined, revised, ruled out, or overturned between admissions. Includes increasing diagnostic specificity (e.g., infection → *Klebsiella pneumonia*) and diagnostic re-evaluation.

2. D2 Therapeutic Trajectory and Optimization

Changes in treatment strategy, medications, or interventions across admissions, including drug addition/removal, dose adjustment, escalation/de-escalation, and management changes.

3. D3 Biomarker and Physiological Trend Reasoning

Changes in laboratory values, vitals, or imaging findings across admissions. Aggregate related biomarkers into a single delta (e.g., CBC, renal panel, liver panel, electrolytes). Maximum of 5 biomarker deltas per admission pair.

4. D4 Causal and Etiological Reasoning

Emit only when source text explicitly expresses causal relationships between admissions using language such as due to, precipitated by, or as a complication of. Do **not** infer causality from temporal proximity.

5. D5 Complication and Adverse Event Tracking

A new complication, adverse event, or clinical problem emerges in admission 2 and was absent in admission 1 (e.g., hospital-acquired infection, organ failure, drug adverse event).

6. D6 Prognostic and Disposition Reasoning

Outcome shifts across admissions, including disposition changes, functional decline, mortality, or pending workup resolution.

Critical Rules

- If admission 2 explicitly mentions passed away, expired, deceased, or died, you **must** emit a D6 Prognostic and Disposition Reasoning delta.
- If admission 1 does not mention a clinical state, set from_state = "not documented".
- For D4_CAUSAL_LINK, only emit when causal language explicitly links admission 1 to admission 2.
- Quote or closely paraphrase source text for from_state and to_state.

Respond strictly in JSON format

```
{
  "deltas": [{
    "type": "<D1... | D2... | D3... | D4... | D5... | D6...>",
    "from_state": "<source or not documented>",
    "to_state": "<source or not documented>",
    "evidence": "<one-sentence factual description>"
  }]
}
```

Figure 6: Full prompt used for cross-admission clinical change detection.

Prompt for Canonical Clinical Reasoning Question Generation

You are a clinical doctor designing assessment questions for senior medical residents.

=== PATIENT CASE (relevant excerpts) ===

{notes_text}

=== IDENTIFIED CHANGES ({dim_name}) ===

{changes_text}

Generate exactly {n_questions} canonical clinical reasoning question(s).

Each question MUST require reasoning across **MULTIPLE notes/time points**.

RECOMMENDED QUESTION TYPES for this dimension

(draw from these as inspiration):

{question_types}

For each question, output:

- clinical_scenario: 2-3 sentence cross-note synthesis capturing the essential clinical trajectory
- core_question: precise open-ended reasoning question
- correct_answer: clinically correct answer with brief explanation
- reasoning_chain: 2-4 reasoning steps
- supporting_deltas: list of delta indices (0-based) supporting this question
- common_errors: 2-3 plausible but incorrect answers reflecting real clinical misconceptions
- question_type: one of {question_types}

Output Format

Respond strictly as a JSON object with a top-level key:

"questions"

containing an array of generated questions.

No markdown code fences.

Figure 7: Representative prompt used in Stage II for canonical question generation.

Prompt for Adversarial Question Generation

Convert to adversarial question.

Clinical scenario:
{scenario}

Core question:
{question}

Correct reasoning:
{reasoning_chain}

Common errors:
{common_errors}

Evidence-based error patterns:
{error_patterns}

DETECTED SIGNAL:
{detected_signal}

You MUST construct a trap matching the detected signal above. Do NOT use a different trap type.

Trap Types

A. TEMPORAL TRAP
The premise subtly assumes that evidence from a **single time point** is sufficient, when correct reasoning requires integration across **multiple admissions/notes**. Embed the error naturally.

B. FALSE CONTINUITY TRAP
The premise subtly assumes that similar findings across time represent the **same continuing event**, although they are actually distinct events with different etiologies.

C. COGNITIVE HEURISTIC TRAP (fallback)
Embed a subtle cognitive bias (e.g., anchoring, premature closure, availability bias) in the premise or expected reasoning path.

CRITICAL RULES

1. The question MUST resemble a normal clinical reasoning question.
2. Do NOT include meta-linguistic phrases such as:
"Identify the error", "Critique the reasoning", or "What is wrong with".
3. The subtle error must be embedded in the premise or expected reasoning path.
4. The correct answer should provide valid clinical reasoning and naturally correct the embedded error.
5. The question should feel like a real clinical assessment rather than a "find-the-mistake" task.

Output JSON

```
{
  "question_text": "...",
  "correct_answer": "...",
  "explanation": "...",
  "trap_type": "Temporal / False_Continuity / Heuristic",
  "trap_explanation": "..."
}
```

No markdown fences.

Figure 8: Representative prompt used to generate adversarial questions in Stage II.

Prompt for MCQ-Unanswerable Generation

Convert to MCQ-Unanswerable.

Clinical scenario:

{scenario}

Core question:

{question}

Correct answer:

{correct_answer}

Common errors:

{common_errors}

Design a clinically reasonable multiple-choice question that **cannot be answered** from the available scenario because critical information is missing.

The question should appear answerable at first glance.

Construction Types

TYPE A – Missing Objective Fact

A specific clinical fact is absent from all notes (e.g., lab value, imaging finding, microorganism, dosage, exact timing).

TYPE B – Broken Cross-Note Inference

Answering requires integrating evidence across admissions, but a critical linking component is missing or ambiguous.

CRITICAL RULES

1. Provide exactly **4 options (A/B/C/D)**.
2. The correct option **MUST** explicitly indicate insufficient information, e.g.,
"Cannot be determined from the clinical notes" or
"The information in the notes is insufficient".
3. The remaining three options must be clinically plausible but unsupported inferences.
4. Do NOT ask rationale or "why" questions.
5. Do NOT use "All of the above" or "None of the above".
6. The explanation **MUST** specify:
(i) whether the question is Type A or Type B,
(ii) what information is missing, and
(iii) why the scenario is unanswerable.
7. The question text **MUST NOT** reveal the answer or explanation.

Output JSON

```
{  
  "question_text": "...",  
  "choices": {"A": "...", "B": "...", "C": "...", "D": "..."},  
  "correct_answer": "A/B/C/D",  
  "explanation": "..."  
}
```

No markdown fences.

Figure 9: Representative prompt used to generate unanswerable multiple-choice questions (MCQ-U) in Stage II.

Open-Ended Question (OEQ) Quality Rubric

1. Answerability (0–3)

Evaluate whether the question can be answered using the available clinical notes.

- **3** = Fully answerable; all necessary evidence explicitly present.
- **2** = Mostly answerable; only minor reasonable inference beyond notes.
- **1** = Partially answerable; substantial inference or assumptions required.
- **0** = Question cannot be answered from the notes; critical information missing.

2. Evidence Consistency (0–3)

Evaluate whether key assertions are supported by note evidence.

- **3** = All key assertions have explicit and direct support in the notes.
- **2** = Most assertions supported; occasional minor inference extends slightly beyond evidence.
- **1** = Some assertions weakly supported or mildly contradictory.
- **0** = Key assertions contradict evidence in the notes.

3. Grounding Fidelity (0–3)

Evaluate whether the answer remains grounded in the clinical notes.

- **3** = Strictly grounded in explicit note evidence; no external knowledge.
- **2** = Largely note-grounded with only minor inference.
- **1** = Introduces some external medical knowledge while retaining a note-based conclusion.
- **0** = Relies substantially on unsupported external knowledge.

4. Reasoning Validity (0–3)

Evaluate the logical correctness of clinical reasoning.

- **3** = Rigorous reasoning demonstrating advanced analytical depth.
- **2** = Logically correct; evidence-to-conclusion steps are valid.
- **1** = Mostly sound but contains logical jumps or unsupported assumptions.
- **0** = Major logical errors or invalid reasoning.

5. Clinical Correctness (0–3)

Evaluate factual medical correctness.

- **3** = Clinically precise, complete, and aligned with best evidence.
- **2** = Clinically correct and consistent with medical consensus.
- **1** = Minor clinical inaccuracies without affecting the core conclusion.
- **0** = Clinically incorrect or misleading.

6. Longitudinal Integration (0–3)

Evaluate whether reasoning integrates temporally distributed evidence.

- **3** = Integrates longitudinal evidence and explains causal mechanisms or clinical significance over time.
- **2** = Correctly integrates multiple time points and tracks clinical changes.
- **1** = Mentions multiple time points without meaningful integration.
- **0** = Relies on a single time point only.

Figure 10: Quality rubric for evaluating open-ended clinical reasoning questions (OEQ).

Adversarial Question (ADVQ) Quality Rubric

1. **Trap Efficacy (0–3)**

Evaluate whether the embedded trap effectively misleads inattentive reasoning.

- **3** = Extremely subtle; even experienced clinicians may fall for it (e.g., anchoring or availability bias).
- **2** = Well-designed; most inattentive reasoners would likely be misled.
- **1** = Partially effective but feels artificial or forced.
- **0** = Trap is obvious and easily detectable with minimal attention.

2. **Subtlety (0–3)**

Evaluate how naturally the embedded error is integrated into the question.

- **3** = Error is effectively invisible without deep analytical reasoning.
- **2** = Naturally embedded and resembles a genuine clinical question.
- **1** = Error is noticeable but partially integrated into the premise.
- **0** = Embedded error is blatant and the question feels artificial.

3. **Clinical Reasoning Soundness (0–3)**

Evaluate the clinical validity of the correct reasoning and conclusion.

- **3** = Rigorous, precise reasoning reflecting senior-level clinical problem solving.
- **2** = Clinically correct and complete; conclusions well supported by the notes.
- **1** = Mostly sound but includes clinical gaps or unsupported assumptions.
- **0** = Contains clinical reasoning errors or reaches an incorrect conclusion.

4. **Cognitive Challenge (0–3)**

Evaluate the level of metacognitive monitoring and analytical reasoning required.

- **3** = Requires systematic cross-note analysis and strong metacognitive monitoring.
- **2** = Requires active inhibition of intuitive reasoning and engagement of analytical reasoning.
- **1** = Requires minor attention but remains largely intuitive.
- **0** = Answerable through simple recall without reflective reasoning.

5. **Clinical Plausibility (0–3)**

Evaluate whether the trap and clinical scenario are realistic.

- **3** = Highly realistic and reflects a documented clinical pitfall.
- **2** = Fully plausible and could realistically occur in practice.
- **1** = Partially plausible but clinically unnatural in some aspects.
- **0** = Clinically implausible or inconsistent with medical facts.

6. **Distinguishability (0–3)**

Evaluate the ability to distinguish different competence levels.

- **3** = Precisely separates experts from novices, including confident-but-wrong vs cautious-but-correct reasoning.
- **2** = Differentiates intermediate from advanced performers.
- **1** = Distinguishes only obviously inexperienced from moderately experienced reasoners.
- **0** = Fails to distinguish competence differences.

Figure 11: Quality rubric for evaluating adversarial clinical reasoning questions (ADVQ).

MCQ-Unanswerable (MCQ-U) Quality Rubric

1. Unanswerability Validity (0–3)

Evaluate whether the question is genuinely unanswerable despite appearing answerable.

- **3** = Appears answerable with a plausible inference path but is genuinely unanswerable.
- **2** = Clearly unanswerable; critical information is genuinely absent.
- **1** = Vaguely unanswerable; careful reasoning may partially infer an answer.
- **0** = Actually answerable from the notes (false unanswerable).

2. Distractor Plausibility (0–3)

Evaluate the realism and misleading potential of incorrect options.

- **3** = Incorrect options are highly plausible and reflect realistic clinical inference traps.
- **2** = All incorrect options appear plausible and require careful elimination.
- **1** = Some distractors are plausible but contain noticeable flaws.
- **0** = Distractors are obviously incorrect and easily eliminated.

3. Premature Closure Risk (0–3)

Evaluate the risk of harmful over-claiming under incomplete evidence.

- **3** = Forced answering is dangerous and may cause serious diagnostic error.
- **2** = Forced answering could lead to clinically significant misjudgment.
- **1** = Forced answering may cause only minor deviation in reasoning or management.
- **0** = Information gap is trivial and unlikely to affect decisions.

4. Cognitive Challenge (0–3)

Evaluate the analytical effort required to confirm uncertainty.

- **3** = Requires deep verification that no information source in the notes can answer the question.
- **2** = Requires careful analysis to confirm that critical information is truly missing.
- **1** = Requires limited attention and mainly recognition of absent information.
- **0** = “Cannot be determined” is immediately obvious.

5. Clinical Scenario Realism (0–3)

Evaluate whether uncertainty and missing information reflect authentic clinical practice.

- **3** = Highly realistic and reflects clinically important uncertainty affecting management.
- **2** = Fully realistic; missing information reflects genuine clinical care.
- **1** = Partially realistic but contains clinically unnatural elements.
- **0** = Unrealistic or inconsistent with routine medical practice.

6. Distinguishability (0–3)

Evaluate the ability to separate cautious reasoners from over-confident responders.

- **3** = Precisely distinguishes over-reaching from uncertainty-aware reasoning.
- **2** = Differentiates intermediate and advanced performers, where stronger reasoners appropriately abstain.
- **1** = Distinguishes only obviously overconfident responders.
- **0** = Fails to distinguish competence differences.

Figure 12: Quality rubric for evaluating unanswerable multiple-choice clinical reasoning questions (MCQ-U).