TemporalEHR-QA: A Benchmark for Temporally Grounded Question Answering over Longitudinal Clinical Notes

Anonymous ACL submission

Abstract

Electronic health records (EHRs) contain temporally evolving clinical narratives. Yet existing clinical question answering (QA) datasets mainly focus on structured queries, isolated notes, or short-context reasoning. We introduce **TemporalEHR-QA**, a benchmark for temporally grounded QA over longitudinal clinical notes collected at a healthcare partner institution. Our dataset construction framework converts raw notes into chronological patient timelines. It then generates schema-constrained temporal QA pairs, and applies two-stage verification using temporal consistency checks and exact evidence matching. The resulting **TemporalEHR-QA** dataset supports diverse temporal reasoning tasks, including progression, recurrence, first/last occurrence, trend analysis, cross-visit comparison, and treatment response. Quality evaluation with both LLM-as-a-Judge and human evaluations ensures strong evidence grounding and temporal consistency. Baseline experiments show that temporally grounded QA over longitudinal clinical histories remains highly challenging for both retrieval and answer generation. This highlights the need for more faithful, evidence-aware, and temporally robust clinical QA systems.

1 Introduction

Electronic health records (EHRs) contain temporally evolving clinical narratives distributed across admissions, outpatient encounters, investigations, and treatment trajectories. Although public resources such as MIMIC-III (Johnson et al., 2016), MIMIC-IV (Johnson et al., 2023), and PhysioNet (Goldberger et al., 2000) have substantially advanced clinical natural language processing (NLP), question answering (QA) over longitudinal clinical notes remains challenging because evidence is temporally dispersed, heterogeneous, and often distributed across long patient histories (Bardhan et al., 2024).

Patient Timeline (Total 3 visits)
Visit 1 – 2019-03-14
"Metformin 500mg daily started for newly diagnosed type 2 diabetes."
Visit 2 – 2019-08-02
"Keep using Metformin."
Visit 3 – 2020-01-18
"Stop using Metformin due to improvement."
Question:
When was Metformin first documented for the patient?
Gold Evidence:
Visit 1 – "Metformin 500mg daily started for newly diagnosed type 2 diabetes."
Answer:
Metformin was **first** documented during Visit 1 on 2019-03-14.
Reasoning Type: First/Last Occurrence

Figure 1: Example of temporally grounded clinical question answering over longitudinal EHR narratives.

Existing clinical QA datasets mainly focus on structured queries, template-derived questions, or short-context reasoning. For example, emrQA (Pampari et al., 2018) studies template-generated QA over clinical notes, while MIMIC-SQL and EHRSQL investigate text-to-SQL reasoning over structured EHR databases (Wang et al., 2020; Lee et al., 2022). Other resources explore graph-based, discharge-summary, radiology, or medication-oriented QA (Park et al., 2021; Raghavan et al., 2021; Bardhan et al., 2022; Lehman, 2022; Soni et al., 2022; Moon et al., 2023). However, these benchmarks generally do not emphasize temporal reasoning over longitudinal patient histories.

Temporal reasoning is fundamental to clinical practice. Clinicians often need to determine when a medication was first documented, whether symptoms improved after treatment, whether a condition recurred, or how patient status changed across visits. Such questions require integrating evidence distributed across temporally separated encounters rather than extracting a single span from one document. Figure 1 illustrates a representative example of temporally grounded QA over longitudinal patient timelines.

Dataset	Longitudinal Context	Temporal Reasoning	Multi-visit Reasoning	Evidence Grounding	Unstructured Notes	Clinical Domain
emrQA (Pampari et al., 2018)	Partial	Limited	Limited	Weak	✓	General EHR
MIMICSQL (Wang et al., 2020)	✗	Structured	✗	–	✗	ICU EHR
EHRSQL (Lee et al., 2022)	✗	Structured	✗	–	✗	EHR Database
DiSCQ (Lehman, 2022)	Limited	Limited	✗	Partial	✓	Discharge
RadQA (Soni et al., 2022)	✗	Minimal	✗	Partial	✓	Radiology
RxWhyQA (Moon et al., 2023)	Partial	Medication	Limited	Partial	✓	Medication
TemporalEHR-QA	✓	**Strong**	✓	**Exact-match**	✓	**Longitudinal Clinical Notes**

Table 1: Comparison between existing clinical QA datasets and TemporalEHR-QA across longitudinal coverage, temporal reasoning, evidence grounding, and clinical-domain characteristics.

To address this gap, we introduce **TemporalEHR-QA**, a benchmark for temporally grounded question answering over longitudinal clinical notes collected at a healthcare partner institution. Our framework converts raw clinical-note rows into chronological patient timelines, generates schema-constrained temporal QA pairs, and applies a two-stage verification pipeline using temporal filtering and exact evidence matching. The resulting dataset preserves visit chronology, timestamps, evidence provenance, and temporal reasoning labels to support evaluation of retrieval, temporal grounding, and faithful longitudinal reasoning.

Our contributions are as follows:

- We introduce TemporalEHR-QA, a benchmark for temporally grounded question answering over longitudinal unstructured clinical notes, covering diverse clinical temporal reasoning categories such as progression, recurrence, trend analysis, first/last occurrence, cross-visit comparison, and treatment response.
- We validate the dataset through both LLM-based and human clinical-domain evaluation, and provide baseline experiments demonstrating that temporally grounded QA over longitudinal clinical histories remains highly challenging for both retrieval and answer generation.

2 Related Work

2.1 Clinical QA and Temporal EHR Benchmarks

Clinical question answering over electronic health records has evolved from short-context span extraction toward increasingly complex forms of clinical reasoning. Early resources such as emrQA (Pampari et al., 2018) focused on template-generated QA over clinical notes, while MIMICSQL and EHRSQL (Wang et al., 2020; Lee et al., 2022) framed clinical QA as text-to-SQL reasoning over structured EHR databases. Other datasets explored specialized settings including discharge-summary QA, radiology QA, and medication-oriented reasoning (Lehman, 2022; Soni et al., 2022; Moon et al., 2023). More recent benchmarks such as TIMER-Bench and TIMER-Instruct (Cui et al., 2025) further emphasize the importance of temporal reasoning in clinical NLP.

Despite these advances, most existing resources focus on isolated notes, structured records, or relatively short contexts, with limited emphasis on reasoning over temporally distributed evidence across long patient histories. In realistic clinical workflows, however, clinicians frequently need to track symptom evolution, treatment response, recurrence, and disease progression across multiple visits. TemporalEHR-QA addresses this gap by focusing on temporally grounded QA over longitudinal clinical-note timelines while preserving explicit visit chronology, evidence provenance, temporal reasoning labels, and exact evidence matching. Table 1 compares TemporalEHR-QA with existing clinical QA benchmarks across longitudinal context, temporal reasoning, evidence grounding, and retrieval characteristics.

2.2 Retrieval and Long-Context Clinical Reasoning

Recent work increasingly highlights retrieval and temporal reasoning as major challenges for longitudinal EHR understanding. Retrieval-oriented approaches such as DR.EHR (Zhao et al., 2025), EHR-RAG (Cao et al., 2026), KG-Retriever (Chen et al., 2024), and GRAIL (Qu and F"arber, 2026) attempt to improve evidence localization across long patient histories, while frameworks such as TRACE (Qu and Färber, 2026) and FHIRPath-

QA (Frew et al., 2026) explore temporal memory and structured reasoning over clinical timelines.

At the same time, long-context clinical language models, including Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), Clinical-Longformer (Li et al., 2022), BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), and GatorTron (Yang et al., 2022), have expanded the amount of clinical text that can be processed within a single context window. However, longer context windows alone does not guarantee temporally faithful reasoning or reliable evidence grounding across heterogeneous patient histories.

TemporalEHR-QA is designed to support the evaluation of these capabilities by combining temporally grounded QA, explicit evidence attribution, retrieval evaluation, and longitudinal temporal reasoning within real clinical-note timelines.

3 Source Corpus and Timeline Construction

3.1 Raw Clinical Notes

The source corpus consists of 191,749 clinical-note rows collected from a healthcare partner institution and associated with 961 unique patients. Each row contains six primary fields: `Visit_DateTime`, `URNumber`, `doc_id`, `doc_text`, `person_id`, and an index column. The notes span multiple years and represent longitudinal documentation across heterogeneous care trajectories. Unlike public critical-care resources such as MIMIC-III and MIMIC-IV (Johnson et al., 2016, 2023), this corpus is drawn from a local health-system setting and is used under institutional governance. The goal is not to release raw clinical notes, but to construct and evaluate privacy-conscious temporal QA annotations grounded in longitudinal clinical documentation.

Only one `doc_text` value is missing, and no exact duplicate rows are identified during preprocessing. Table 2 summarizes the corpus after preprocessing and visit-level aggregation. Additional corpus plots in Appendix A show the temporal coverage of clinical-note timestamps, the right-skewed distribution of note lengths, and the long-tailed distribution of visits per patient.

3.2 Text Normalization and Visit Aggregation

We convert `Visit_DateTime` from millisecond Unix timestamps into standard datetime representations and apply light normalization to the clinical text. The preprocessing pipeline removes repeated whitespace and newline artifacts while preserving punctuation-sensitive clinical expressions. We retain the original surface form as much as possible because punctuation, abbreviations, and formatting patterns may carry clinically relevant temporal information. We discard empty cleaned notes before timeline construction.

Statistic	Value
Raw note rows	191,749
Unique patients	961
Visit-level rows after aggregation	170,847
Eligible patients with 2–1000 visits	902
Patients with 2–10 visits	107
Patients with 11–100 visits	466
Patients with 101–1000 visits	329
Single-visit patients excluded	30
Patients with >1000 visits excluded	29
Maximum visits for one patient	3,516

Table 2: Corpus statistics after preprocessing and visit-level aggregation.

Clinical documentation frequently contains multiple note fragments recorded at the same timestamp for the same patient. To reduce fragmentation and preserve coherent temporal events, we group rows by `person_id` and `Visit_DateTime`. Notes within the same group are concatenated into a visit-level context while retaining document identifiers as provenance metadata.

After aggregation, patient visits are sorted chronologically and serialized into structured patient timelines:

```
{person_id:          int,       visits:
[{visit_datetime:   str,  text:   str},
...]}
```

Each line in `patient_sequences.jsonl` therefore represents one temporally ordered patient history.

3.3 Patient Eligibility and Visit Groups

The number of visits per patient varies dramatically. Short histories can often be processed in full, while long histories require sampling and chunking strategies to preserve temporal coverage within practical context limits. We retain patients with 2–1000 visits. Single-visit patients are excluded because they cannot support cross-visit temporal reasoning, and patients with more than 1000 visits are excluded from this version because they require specialized extreme-long-context summarization. Retained patients are assigned to three groups: 2–10, 11–100, and 101–1000 visits.

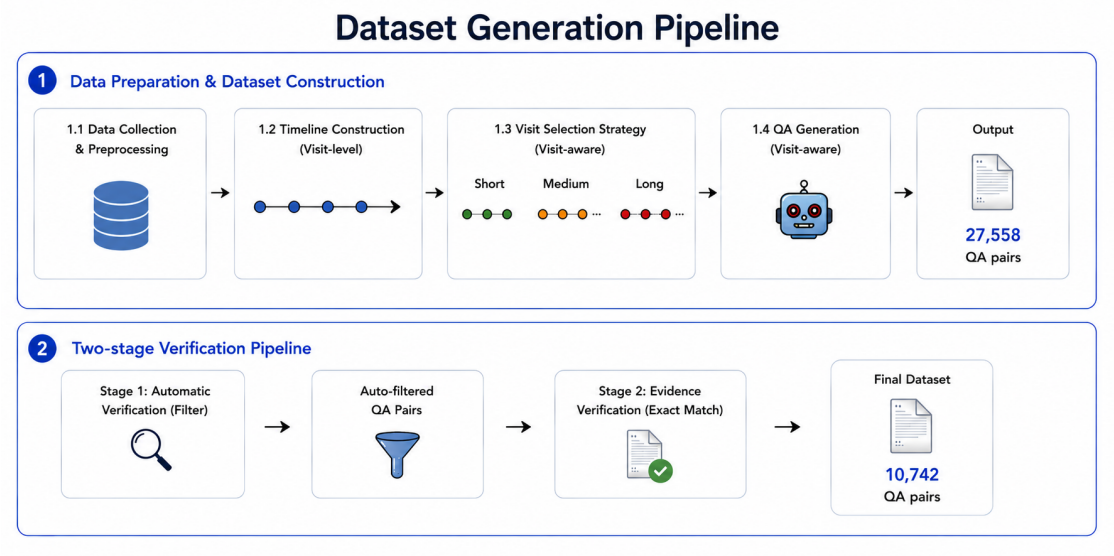

Figure 2: Overall framework for constructing the temporally grounded longitudinal clinical QA dataset. The pipeline consists of preprocessing and visit-level timeline construction, visit-aware QA generation, two-stage temporal and evidence verification, and final dataset analysis.

4 Dataset Construction Method

Figure 2 presents the overall dataset construction framework. The pipeline converts raw clinical-note rows into visit-level timelines, generates temporally grounded QA pairs using visit-aware sampling, and applies a two-stage verification process to ensure evidence fidelity.

4.1 Visit-Aware Timeline Sampling

The dataset was generated in ten batches, each using up to 100 preprocessed patients while preserving coverage across the three visit groups. The pipeline uses a checkpoint file, `processed_patient_ids.txt`, to avoid reprocessing patients between batches.

The sampling strategy varies according to timeline length to balance temporal coverage and prompt tractability. For patients with 2–10 visits, the full timeline is included. For medium-length timelines (11–100 visits), all visits are retained when the patient has at most 30 visits. Otherwise, a U-shaped sampling strategy selects the first 12 and last 18 visits to preserve both early-history and recent clinical evidence, where clinically important information is often concentrated (Cao et al., 2026).

For long timelines containing 101–1000 visits, the pipeline combines a global view with local chunk-based views. The global view captures coarse-grained longitudinal context by selecting the first 15 and last 15 visits, while the local views use overlapping chunks of 30 visits with overlap 5 to preserve fine-grained temporal continuity between adjacent clinical events. This hierarchical combination of global context and local chunking is motivated by prior work on long-document retrieval and hierarchical reasoning, suggesting that combining global temporal structure with fine-grained local context can improve reasoning over long patient timelines (Choi et al., 2016; Liu et al., 2021).

This design aims to preserve clinically informative temporal coverage while keeping generation prompts tractable. The global view supports questions about the beginning and end of a trajectory; local chunks support dense temporal reasoning over adjacent clinical events.

4.2 Schema-Constrained QA Generation

Candidate QA pairs are generated using GPT-5.4-nano (OpenAI, 2026) through the OpenAI Responses API with strict JSON-schema output constraints. We selected GPT-5.4-nano (OpenAI, 2026) because the task primarily requires high-throughput structured generation, temporal evidence extraction, and schema-constrained annotation rather than open-ended clinical decision making. Recent studies have shown that smaller instruction-tuned language models can produce reliable synthetic QA annotations when generation is constrained through structured prompting, verification pipelines, and evidence-grounded filtering (Pampari et al., 2018; Cui et al., 2025). In our

Field	Description
person_id	Internal patient identifier used for timeline construction and evidence tracing. Removed or anonymized before external release.
num_visits_in_prompt	Number of visits included in the sampled timeline or chunk provided to the generation model.
visit_group	Visit-frequency category: 2–10, 11–100, or 101–1000 visits.
timeline_sampling_strategy	Timeline selection strategy used during QA generation, including full timeline, U-shaped sampling, global sampling, or overlapping local chunking.
question	Natural-language clinical question generated from the longitudinal patient timeline.
answer	Evidence-grounded answer supported by cited clinical-note snippets.
reasoning_type	Temporal reasoning category, including before–after, progression, first/last occurrence, recurrence count, trend analysis, comparison across visits, treatment response, and other.
difficulty	Difficulty level of the QA pair: easy, medium, or hard.
evidence	Supporting evidence snippets linked to sampled visit indices, original visit indices, timestamps, and source text spans.
temporal_signal	Short free-text description of the temporal relationship underlying the question.

Table 3: Core schema fields in TemporalEHR-QA. The schema preserves temporal ordering, evidence provenance, and longitudinal reasoning annotations for temporally grounded clinical QA.

setting, the downstream two-stage verification process further reduces dependence on unconstrained model generation by enforcing temporal consistency and exact evidence matching against the original clinical notes. The complete prompt template, JSON schema are provided in Appendix B.

5 Dataset Verification and Analysis

Following schema-constrained generation, all candidate QA pairs undergo a two-stage verification pipeline designed to improve temporal validity and evidence fidelity. The first stage applies deterministic temporal and structural filtering to remove malformed QA pairs and enforce temporally consistent multi-visit reasoning, following prior work showing the importance of structured filtering and schema-guided QA validation in synthetic clinical QA generation (Bai et al., 2024). The second stage performs exact evidence matching against the original clinical-note text to verify evidence provenance and supporting context, motivated by evidence-grounded QA and supporting-fact verification frameworks that emphasize faithful evidence attribution for multi-hop reasoning tasks (Yang et al., 2018).

5.1 Dataset Schema

Each final JSONL record represents one patient timeline together with its associated QA pairs. The schema stores patient-level metadata describing the sampled timeline and a list of generated QA items. Every QA item stores semantic and temporal annotations, including reasoning type, difficulty level, evidence snippets, and temporal signals. Evidence is linked back to original visits using visit indices and timestamps so that answer grounding can be verified against the source notes. Table 3 summarizes the core schema fields.

5.2 Round 1: Automatic Temporal and Evidence Filtering

The first verification stage normalizes generated outputs by attaching original visit indices and timestamps to all evidence snippets, followed by deterministic temporal and structural filtering. QA pairs are removed if they contain malformed questions, empty answers, missing evidence, invalid reasoning types, temporally inconsistent evidence ordering, unsupported clinical inferences, or duplicate question–answer content. For temporal reasoning categories such as progression, trend, recurrence, comparison across visits, and treatment response, the pipeline additionally requires evidence from multiple temporally distinct visits.

This stage is intentionally conservative to improve temporal validity and evidence grounding. Across ten generation batches, the model initially produced 27,558 candidate QA pairs. After Round 1 filtering, 18,867 QA pairs were retained, corresponding to a retention rate of 68.38%.

5.3 Round 2: Exact Evidence Matching

The second verification stage performs exact evidence matching against the original clinical-note text. For each evidence item, the pipeline retrieves the corresponding visit using the patient identifier, visit timestamp, and original visit index, then checks whether the cited evidence snippet can be

directly matched to the source note. A QA pair is retained only when all evidence snippets are successfully aligned with the original clinical documentation.

This stage provides a conservative evidence-provenance check that prioritizes faithful grounding and traceability to the source notes. The matching algorithm additionally supports snippets containing ellipses by splitting them into shorter segments and verifying each segment independently against the original text. Across the merged ten-batch output, 10,742 QA pairs retained complete evidence alignment, while 8,125 candidate QA pairs were removed by the strict evidence-matching criterion. Although this conservative filtering improves evidence fidelity, it may exclude clinically valid paraphrastic or semantically equivalent evidence.

5.4 Human and LLM-Based Quality Evaluation

To estimate residual annotation noise after automatic verification, we conducted a structured quality evaluation on a stratified sample of 1,000 QA pairs covering different reasoning types, difficulty levels, and visit-frequency groups. Human evaluation by a clinical-domain researcher was used as the primary quality assessment, while GPT-5.5 with Thinking Mode acting as an LLM-as-a-judge was used as a supplementary consistency check. For transparency and reproducibility, Appendix C.1 reports the full instructions given to the human evaluator, including the annotation criteria, privacy precautions, risk disclosure, and a schematic example of the annotation interface.

Each sampled QA pair was evaluated along five dimensions: evidence match, question validity, answer correctness, evidence grounding, and temporal consistency.

Table 4 summarizes the evaluation results. Human evaluation indicates consistently high dataset quality across all five evaluation dimensions, including 98.5% evidence match, 97.3% question validity, 95.2% answer correctness, 96.1% evidence grounding, and 96.8% temporal consistency. The LLM-based judge yielded generally more conservative assessments, especially for answer correctness and evidence grounding. Taken together, these findings suggest that the proposed verification pipeline effectively reduces residual annotation noise while maintaining strong temporal consistency and evidence fidelity.

Metric	LLM-T	LLM-F	Human-T	Human-F
Evidence match	962	38	985	15
Question validity	934	66	973	27
Answer correctness	881	119	952	48
Evidence grounding	902	98	961	39
Temporal consistency	914	86	968	32

Table 4: Quality evaluation on 1,000 QA pairs. LLM-T/LLM-F denote true/false judgments from GPT-5.5 with Thinking Mode prompted as a clinical expert. Human-T/Human-F denote judgments from a clinical-domain researcher.

The assessed sample intentionally emphasized challenging long-history cases, with most QA pairs drawn from patients containing 101–1000 visits. Detailed evaluation distributions are provided in Appendix C.2.

5.5 Final Dataset Statistics

The final dataset contains 10,742 evidence-matched QA pairs across 829 patients. Appendix D summarize the distributions of difficulty levels, temporal reasoning types, and patient visit-frequency groups. Medium-difficulty questions constitute the majority of the dataset, reflecting the need to integrate multiple evidence snippets while remaining explicitly grounded in the clinical notes. Comparison, progression, and before–after reasoning are the most common categories, highlighting the dataset's emphasis on longitudinal evidence integration rather than isolated span extraction.

The dataset also contains substantial coverage of medium- and long-history patient timelines. While some QA pairs rely on evidence from nearby visits within a short hospitalization period, many require information distributed across months or years of patient history. This variation reflects the temporal complexity of real-world clinical reasoning, which often involves both short-term events and long-term disease progression or treatment trajectories.

6 Baseline Evaluation

The primary goal of the experiments is to characterize the difficulty of temporally grounded question answering over longitudinal clinical notes rather than to exhaustively optimize model performance. We therefore evaluate representative retrieval and reader architectures under both end-to-end and oracle-evidence settings. This design allows us to separate two major sources of difficulty: (1) locating the correct temporally distributed evidence within long patient timelines, and (2) synthesizing

Retriever	R@5	R@10	R@20	Cov@5	Cov@10	Cov@20
BM25	11.9	14.8	18.9	8.6	11.3	14.2
MedCPT	12.2	15.9	**20.4**	8.4	11.7	15.0
NV-Embed-v2	12.2	15.3	19.7	8.6	11.5	15.0
Hybrid BM25+MedCPT	**13.1**	**16.3**	20.2	**9.6**	**12.5**	**15.6**

Table 5: Retrieval performance across different retrieval depths. Recall@K indicates whether at least one gold evidence visit is retrieved, while Exact Coverage@K requires retrieval of all gold evidence visits associated with a QA pair.

temporally coherent answers from retrieved evidence.

6.1 Experimental Setup

We evaluate BM25 (Robertson and Walker, 1994), MedCPT (Jin et al., 2023), NV-Embed-v2 (Lee et al., 2024), and a hybrid BM25+MedCPT5 (Robertson and Walker, 1994; Jin et al., 2023) retriever on the 10,742 verified QA pairs. For retrieval, the search space is restricted to each patient's longitudinal timeline, where each visit-level note is treated as an independent document.

For question answering, we evaluate ClinicalBERT (Alsentzer et al., 2019), Longformer (Beltagy et al., 2020), and Qwen2.5-7B-Instruct (Qwen Team, 2025) under both end-to-end and oracle-evidence settings. The experiments are designed to evaluate both evidence retrieval over long patient timelines and temporal answer synthesis from longitudinal clinical notes. Additional implementation details, model configurations, hardware specifications, model sizes, and approximate computational budget are provided in Appendix E.

6.2 Retrieval Performance

Table 5 reports retrieval performance across different retrieval depths. Increasing K consistently improves both Recall@K and Exact Coverage@K, confirming that top-5 retrieval alone is insufficient for longitudinal clinical QA. Nevertheless, retrieval remains difficult even at larger retrieval depths. The best Recall@20 reaches only 20.4% with MedCPT (Jin et al., 2023), while the best Exact Coverage@20 is 15.6% using the hybrid BM25+MedCPT (Robertson and Walker, 1994; Jin et al., 2023) retriever.

The hybrid retriever achieves the strongest Exact Coverage across all values of K, suggesting that lexical overlap and clinical semantic similarity provide complementary retrieval signals. However, the relatively small gains between sparse, dense, and hybrid retrieval indicate that the primary challenge lies not only in retriever selection, but in identifying temporally distributed evidence across long and heterogeneous patient histories.

Retrieval difficulty scales sharply with timeline length. While Recall@5 approaches 99.5% for patients with 2–10 visits, performance drops to approximately 8% for patients with more than 100 visits, indicating that long-horizon evidence localization is the dominant retrieval bottleneck in the dataset. Retrieval performance also degrades consistently across increasing difficulty levels and complex temporal reasoning categories such as progression and trend reasoning. Appendix F.1 further breaks retrieval performance down by timeline length, difficulty level, and reasoning type, showing that retrieval becomes progressively harder as temporal reasoning requirements become more complex and patient histories grow longer.

6.3 End-to-End and Oracle-Evidence QA

To separate retrieval errors from reasoning errors, we evaluate QA models under two settings: end-to-end hybrid retrieval and oracle evidence. In the end-to-end setting, readers receive the top retrieved visits from the hybrid retriever, while in the oracle setting, readers receive the gold evidence visits directly.

Table 6 reports the full retriever–reader evaluation matrix. Across all retrieval settings, generative readers consistently outperform extractive models by large margins, while improvements from stronger retrieval models remain comparatively modest. Extractive readers achieve token F1 below 9% across most configurations, suggesting that many questions require temporal synthesis across multiple visits rather than direct span extraction. In contrast, Qwen2.5-7B (Qwen Team, 2025) achieves the strongest performance, reaching 19.63 token F1 and 0.716 BERT-F1 with NV-Embed (Lee et al., 2024) retrieval.

To further separate retrieval errors from reason-

Retriever	Reader	EM	Precision	Recall	Token F1	BERT-F1
BM25	ClinicalBERT	0.10	0.6967	0.6277	6.43	0.6599
	Longformer	0.10	0.7007	0.6327	7.20	0.6645
	Qwen2.5-7B	**0.15**	0.7181	0.6831	16.43	0.6997
MedCPT	ClinicalBERT	0.10	0.7046	0.6339	7.29	0.6669
	Longformer	0.08	0.7098	0.6389	8.33	0.6720
	Qwen2.5-7B	0.11	0.7290	0.6947	18.54	0.7109
NV-Embed	ClinicalBERT	0.09	0.7044	0.6329	7.34	0.6663
	Longformer	0.07	0.7080	0.6369	8.07	0.6701
	Qwen2.5-7B	0.10	**0.7340**	**0.7004**	**19.63**	**0.7163**

Table 6: End-to-end QA performance across retrieval and reader combinations. EM and Token F1 measure lexical overlap, while Precision, Recall, and BERT-F1 are computed using Bio_ClinicalBERT embeddings.

Setting	Reader	Token F1	BERT-F1
Hybrid Retrieval	ClinicalBERT	7.2	0.767
	Longformer	8.1	0.815
	Qwen2.5-7B	**36.3**	**0.886**
Oracle Evidence	ClinicalBERT	3.2	0.660
	Longformer	4.0	0.788
	Qwen2.5-7B	23.9	0.865

Table 7: QA performance under end-to-end hybrid retrieval and oracle-evidence settings. Oracle evidence bypasses retrieval and provides gold evidence visits directly to the reader.

ing errors, we additionally evaluate readers under an oracle-evidence setting. In the end-to-end setting, readers receive the top retrieved visits from the hybrid retriever, while in the oracle setting, readers receive the gold evidence visits directly.

Table 7 summarizes the comparison between hybrid retrieval and oracle evidence. Qwen2.5-7B (Qwen Team, 2025) substantially outperforms extractive readers in both settings, achieving 36.3 token F1 and 0.886 BERT-F1 under hybrid retrieval. Interestingly, performance slightly decreases under oracle evidence, suggesting that chronologically adjacent visits retrieved by the hybrid pipeline may provide useful contextual grounding for temporal reasoning even when they are not part of the gold evidence annotations.

Additional analyses stratified by difficulty level, visit-group size, and reasoning type are provided in Appendix F.2. Overall, the results indicate that evidence retrieval remains a major bottleneck for longitudinal clinical QA, whereas instruction-tuned generative models are substantially better suited for temporally grounded answer synthesis.

7 Conclusion

We introduced TemporalEHR-QA, a benchmark for temporally grounded question answering over longitudinal clinical notes. The dataset is constructed through a pipeline that converts raw EHR note rows into chronological patient timelines, generates schema-constrained temporal QA pairs, and verifies them using temporal filtering and exact evidence matching. The final benchmark contains 10,742 evidence-grounded QA pairs spanning diverse forms of longitudinal clinical reasoning, including progression, recurrence, treatment response, trend analysis, first/last occurrence, and cross-visit comparison. Quality evaluation using both human review and LLM-based assessment indicates strong evidence grounding and temporal consistency. Baseline experiments further show that temporally grounded QA over long patient histories remains highly challenging, particularly for evidence retrieval and long-horizon temporal reasoning. While generative readers substantially outperform extractive architectures, overall performance remains far from solved, highlighting the need for more temporally aware, retrieval-robust, and faithfully grounded clinical QA systems.

Limitations

TemporalEHR-QA is intended to move clinical QA beyond isolated note understanding toward longitudinal temporal reasoning over real patient histories. At the same time, the benchmark reflects several practical trade-offs that arise when constructing temporally grounded datasets from real clinical narratives.

The dataset is derived from a single health-

system environment, meaning that its temporal patterns, note styles, and clinical workflows may not fully represent other institutions or specialties. In addition, although the construction pipeline combines temporal-rule verification, exact evidence matching, and a structured 1,000-sample quality evaluation, the dataset is not fully clinician-adjudicated at scale. Some residual annotation noise likely remains, particularly for clinically nuanced cases involving implicit progression, treatment response, or evidence distributed across distant visits.

Our evidence-matching design also intentionally prioritizes faithful grounding to the source notes. This improves provenance reliability and reduces unsupported answers, but it may also underrepresent clinically valid paraphrases or higher-level summaries that clinicians would naturally infer from the timeline.

The experiments further show that retrieval over long patient histories remains extremely challenging. Performance drops sharply as timelines grow longer, especially for patients with hundreds of visits. While we evaluate sparse, dense, hybrid, and oracle-evidence settings, many promising directions remain unexplored, including hierarchical timeline retrieval, memory-augmented reasoning, temporally aware reranking, and multi-hop retrieval strategies designed specifically for longitudinal clinical records.

Finally, because the benchmark is derived from real clinical notes, privacy-preserving release remains an ongoing challenge. Preserving longitudinal temporal structure while safely removing potentially identifying information requires careful balancing between research utility and patient confidentiality.

Ethics, Privacy, and Release Plan

TemporalEHR-QA is constructed from real longitudinal clinical notes collected under institutional governance at a healthcare partner institution. Because the source documents contain sensitive patient information, we do not release raw clinical notes, original timelines, document identifiers, or patient identifiers.

The source clinical notes were accessed under institutional governance and privacy-review procedures at a university-affiliated healthcare partner for research on privacy-preserving clinical NLP. The research team did not directly recruit patients or obtain new study-specific consent from individual patients as part of constructing this benchmark; instead, data access and use were governed by institutional procedures that determined the permitted research use of the clinical documentation. The data are used only to construct privacy-conscious, temporally grounded QA annotations for clinical NLP research, and the benchmark is not designed for direct clinical deployment or decision-making. To reduce privacy risk, we do not release raw clinical notes, original patient timelines, document identifiers, patient identifiers, or other directly identifying information. Any public or controlled-access release of derived QA annotations remains subject to institutional approval, ethics review where required, privacy review, and applicable data-use agreements.

Our release strategy is designed to support reproducible research while minimizing privacy risk. The planned public release includes the dataset schema, generation prompts, filtering rules, evaluation scripts, aggregate statistics, and de-identified QA annotations, subject to institutional approval and privacy review.

Protecting temporal information is particularly important in longitudinal clinical data. Rather than releasing real timestamps, visits may be represented using shifted dates, relative temporal offsets, visit indices, or coarse-grained temporal intervals. These transformations preserve temporal ordering and longitudinal reasoning structure while reducing re-identification risk.

Because evidence snippets originate from real clinical narratives, they require additional safeguards beyond standard de-identification. All released evidence will undergo automated filtering and manual privacy review. If verbatim release is considered unsafe, evidence may instead be represented using abstracted descriptors or made available only through controlled-access agreements.

We therefore envision two complementary release modes: a public research release containing fully de-identified metadata, prompts, code, aggregate statistics, and synthetic examples; and a controlled-access release of de-identified QA pairs and evidence annotations under a data-use agreement. The benchmark is intended solely for research on temporally grounded clinical QA and is not designed for direct clinical deployment or decision-making use.

References

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

Fan Bai, Keith Harrigian, Joel Stremmel, Hamid Hassanzadeh, Ardavan Saeedi, and Mark Dredze. 2024. Give me some hard questions: Synthetic data generation for clinical qa. *Preprint*, arXiv:2412.04573.

Jayetri Bardhan, Anthony Colas, Kirk Roberts, and Daisy Zhe Wang. 2022. DrugEHRQA: A question answering dataset on structured and unstructured electronic health records for medicine related queries. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1083–1092, Marseille, France. European Language Resources Association.

Jayetri Bardhan, Kirk Roberts, and Daisy Zhe Wang. 2024. Question answering for electronic health records: A scoping review of datasets and models. *Journal of Medical Internet Research*, 26:e53636.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Lang Cao, Qingyu Chen, and Yue Guo. 2026. Ehr-rag: Bridging long-horizon structured electronic health records and large language models via enhanced retrieval-augmented generation. *arXiv preprint arXiv:2601.21340*.

Weijie Chen, Ting Bai, Jinbo Su, Jian Luan, Wei Liu, and Chuan Shi. 2024. Kg-retriever: Efficient knowledge indexing for retrieval-augmented large language models. *arXiv preprint arXiv:2412.05547*.

Eunsol Choi, Daniel Hewlett, Alexandre Lacoste, Illia Polosukhin, Jakob Uszkoreit, and Jonathan Berant. 2016. Hierarchical question answering for long documents. *arXiv preprint arXiv:1611.01839*.

Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H Shah. 2025. Timer: Temporal instruction modeling and evaluation for longitudinal clinical records. *npj Digital Medicine*, 8(1):577.

Michael Frew, Nishit Bheda, and Bryan Tripp. 2026. Fhirpath-qa: Executable question answering over fhir electronic health records. *arXiv preprint arXiv:2602.23479*.

Ary L. Goldberger, Luís A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Qiao Jin, Bhuwan Dhingra, William W. Cohen, and Xinghua Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alyssa Gayles, Ahmad Shammout, Steven Horng, Tom J. Pollard, Shih-Cheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1–13.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, and Bryan Catanzaro. 2024. NV-Embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. 2022. EHRSQL: A practical text-to-SQL benchmark for electronic health records. In *Advances in Neural Information Processing Systems*, volume 35, pages 15589–15601.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Eric Lehman. 2022. Learning to Ask Like a Physician: a Discharge Summary Clinical Questions (DiSCQ) Dataset. *PhysioNet*. Version 1.0.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

Ye Liu, Kazuma Hashimoto, Yingbo Zhou, Semih Yavuz, Caiming Xiong, and Philip Yu. 2021. Dense hierarchical retrieval for open-domain question answering. In *Findings of EMNLP*.

Sungrim Moon, Huan He, Heling Jia, Hongfang Liu, and Jungwei Wilfred Fan. 2023. Extractive clinical question-answering with multianswer and multifocus questions: Data set development and evaluation study. *JMIR AI*, 2(1):e41818.

OpenAI. 2026. Introducing gpt-5.4 mini and nano. `https://openai.com/index/introducing-gpt-5-4-mini-and-nano/`. Accessed: 2026-05-26.

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.

Junwoo Park, Yoon Cho, Haneol Lee, Jaegul Choo, and Edward Choi. 2021. Knowledge graph-based question answering with electronic health records. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pages 36–53. PMLR.

Zhan Qu and Michael F"arber. 2026. Grail: Geometry-aware retrieval-augmented inference with llms over hyperbolic representations of patient trajectories. *arXiv preprint arXiv:2602.12828*.

Zhan Qu and Michael Färber. 2026. Trace: Temporal reasoning via agentic context evolution for streaming electronic health records (ehrs). *arXiv preprint arXiv:2602.12833*.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Preethi Raghavan, Jennifer J. Liang, Diwakar Mahajan, Rachita Chandra, and Peter Szolovits. 2021. emrKBQA: A clinical knowledge-base question answering dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73. Association for Computational Linguistics.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.

Sarvesh Soni, Meghana Gudala, Atieh Pajouhi, and Kirk Roberts. 2022. RadQA: A question answering dataset to improve comprehension of radiology reports. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Ping Wang, Tian Shi, and Chandan K. Reddy. 2020. Text-to-SQL generation for question answering on electronic medical records. In *Proceedings of The Web Conference 2020*, pages 350–361, New York, NY, USA. Association for Computing Machinery.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *Preprint*, arXiv:2203.03540.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of EMNLP*.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Puneet Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297.

Zhengyun Zhao, Huaiyuan Ying, Yue Zhong, and Sheng Yu. 2025. Dr. ehr: Dense retrieval for electronic health record with knowledge injection and synthetic data. *arXiv preprint arXiv:2507.18583*.

This appendix provides additional corpus statistics, dataset analyses, and fine-grained baseline evaluations that complement the main paper.

A Additional Corpus Details

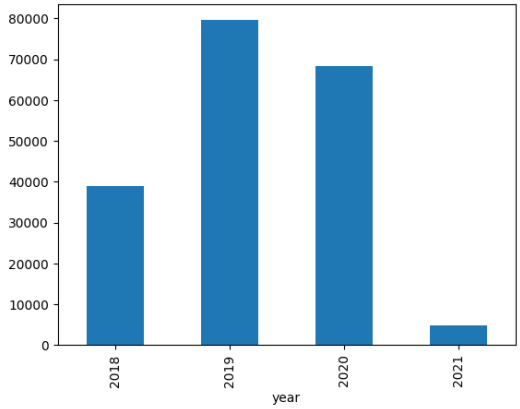

Figure 3: Distribution of clinical-note timestamps by year.

Figure 3 presents the yearly distribution of clinical-note timestamps in the source corpus and illustrates the broad longitudinal temporal coverage of the collected EHR narratives.

Figure 4 reports the distribution of clinical-note lengths after preprocessing and visit-level aggregation, highlighting the substantial variability and right-skewed nature of clinical documentation.

Figure 5 shows the distribution of visit counts per patient and demonstrates the long-tailed structure of longitudinal patient histories, motivating the need for visit-aware sampling strategies.

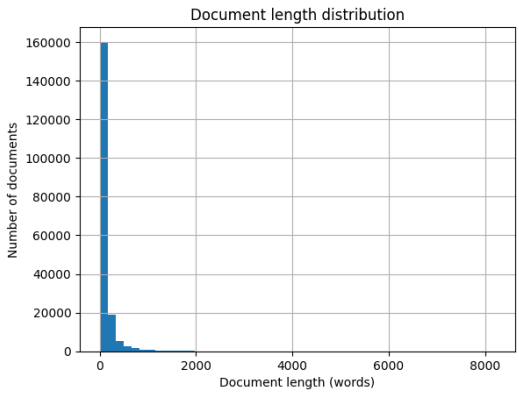

Figure 4: Distribution of clinical-note lengths measured in words.

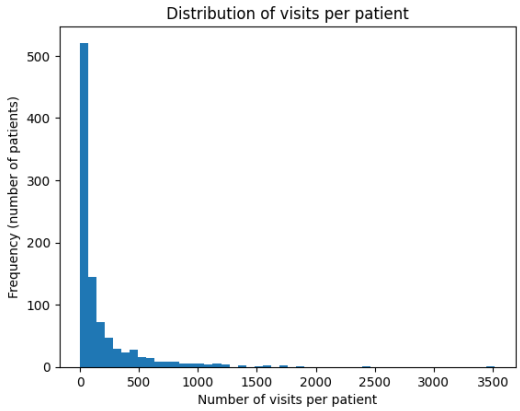

Figure 5: Distribution of visit counts per patient after visit-level aggregation.

B Generation Schema and Prompt Configuration

B.1 System Prompt Summary

The QA generation pipeline uses a system prompt that instructs the model to generate temporally grounded, evidence-supported QA pairs from longitudinal patient timelines. The prompt emphasizes temporal reasoning, evidence faithfulness, privacy preservation, and clinically meaningful question construction.

The model is instructed to:

- generate high-quality non-redundant QA pairs grounded only in the provided timeline;
- prioritize temporal reasoning across visits and timestamps;
- avoid unsupported inference or hallucinated clinical conclusions;
- avoid personally identifying information;
- return outputs strictly in JSON format following the predefined schema.

The prompt also defines the temporal reasoning taxonomy:

- `before_after`
- `progression`
- `first_last_occurrence`
- `recurrence_count`
- `trend`
- `comparison_across_visits`
- `treatment_response`
- `other`

Difficulty levels are defined as:

- **easy**: direct time-aware evidence;
- **medium**: requires linking temporally distinct evidence;
- **hard**: requires multi-visit longitudinal reasoning.

B.2 JSON Schema for QA Generation

Candidate QA pairs are generated under strict JSON-schema constraints using the OpenAI Responses API. The schema enforces structural consistency and evidence attribution during generation.

```
{
  "qas": [
    {
      "question": "string",
      "answer": "string",
      "reasoning_type": {
        "enum": [
          "before_after",
          "progression",
          "first_last_occurrence",
          "recurrence_count",
          "trend",
          "comparison_across_visits",
          "treatment_response",
          "other"
        ]
      },
      "difficulty": {
       "enum": ["easy", "medium", "hard"]
```

```
      },
      "evidence": [
        {
          "visit_index": "integer",
          "evidence_snippet": "string"
        }
      ],
      "is_temporal": "boolean",
      "temporal_signal": "string"
    }
  ]
}
```

Each QA item is required to contain evidence-supported answers linked to specific visits through visit indices and evidence snippets.

C Additional Manual Quality Evaluation

C.1 Quality Evaluation Instructions, and Risk Disclosure

We conducted a structured evaluation to assess the residual noise of the final QA pairs after automatic temporal-rule verification and exact evidence matching. The evaluation was performed on a stratified sample of 1,000 QA pairs. Stratification was based on temporal reasoning type, difficulty level, and visit-frequency group to ensure that the evaluation covered both simple single-visit questions and harder longitudinal questions requiring evidence across multiple visits.

The human quality evaluation was conducted by a clinical-domain researcher. Before beginning the evaluation, the evaluator was given written instructions describing the purpose of the task, the information shown for each QA instance, the annotation criteria, and the privacy precautions required when handling clinical text. For each QA pair, the evaluator was shown the question, gold answer, cited evidence span, visit timestamp or visit index, relevant surrounding clinical-note context, temporal reasoning category, and difficulty label.

Instructions shown to the evaluator. You will evaluate a sample of question–answer pairs generated from longitudinal clinical-note timelines. The purpose of this evaluation is to assess the quality of the QA annotations, not to evaluate patient care, clinician performance, or the medical correctness of the original clinical documentation.

For each item, you will be shown the following fields:

- the generated clinical question;
- the gold answer;
- the cited evidence snippet;
- the visit timestamp or visit index;
- the relevant surrounding clinical-note context;
- the temporal reasoning category; and
- the difficulty label.

Please judge each item using the five binary criteria below. Mark each criterion as TRUE if the item satisfies the criterion, and FALSE otherwise.

1. **Evidence match:** Does the cited evidence snippet appear in the source note and correspond to the referenced visit?
2. **Question validity:** Is the question clinically meaningful, answerable, and not ambiguous given the provided context?
3. **Answer correctness:** Does the answer correctly answer the question?
4. **Evidence grounding:** Is the answer sufficiently supported by the cited evidence and surrounding note context?
5. **Temporal consistency:** Does the answer respect the chronological order of visits, temporal expressions, and cross-visit context?

Please mark an item as FALSE for the relevant criterion if the answer requires unsupported clinical inference, if the cited evidence does not support the answer, if the temporal ordering is inconsistent, or if the question cannot be answered from the provided context. Do not use external medical knowledge to infer answers beyond what is supported by the displayed evidence and note context.

Privacy and confidentiality instructions. The evaluation material may contain sensitive clinical information. Do not copy, download, screenshot, redistribute, or discuss any item outside the evaluation environment. Do not attempt to identify any patient, clinician, institution, or other individual from the clinical text. If you encounter text that appears to contain directly identifying information, unusual identifiers, or sensitive content not required for the evaluation, flag the item for review rather than recording or sharing the content elsewhere.

Risk disclosure. The main risk to the evaluator is exposure to sensitive clinical narratives and potentially distressing medical content. The evaluation task does not require making clinical decisions, contacting patients, or providing medical advice. The evaluator may skip or flag any item that appears inappropriate, overly sensitive, or outside the intended evaluation scope. All evaluation outputs are used only for research quality assessment of the dataset.

Annotation criteria and protocol. Each instance was annotated using five binary validity criteria:

- **Evidence match**: whether the cited evidence span appears in the source note and corresponds to the referenced visit.
- **Question validity**: whether the question is clinically meaningful, answerable, and not ambiguous given the available note context.
- **Answer correctness**: whether the gold answer correctly answers the question.
- **Evidence grounding**: whether the cited evidence sufficiently supports the answer.
- **Temporal consistency**: whether the answer respects the chronological order, temporal expressions, and cross-visit context in the patient timeline.

Human evaluation was used as the primary estimate of dataset quality. In parallel, we also applied an LLM-based judge using the same five criteria. The LLM judge was instructed to mark an item as invalid if the answer was unsupported, temporally inconsistent, or not directly inferable from the cited evidence and note context. We then compared the LLM judgments with human judgments to identify systematic disagreement patterns.

The disagreement analysis shows that the LLM judge was more conservative than the human evaluator. In several cases, the LLM marked answers as FALSE because it failed to interpret clinical temporal shorthand, such as "2/7" meaning two days and "1/52" meaning one week. Other disagreements occurred when the LLM did not fully use information from earlier or later visits, for example when a symptom was documented in one note but resolved in a subsequent note. These cases suggest that human evaluation remains important for assessing temporally grounded clinical QA, especially when answers require domain-specific abbreviation interpretation or longitudinal context integration.

Question: When was the symptom first documented as improving?

Gold answer: The symptom was first documented as improving at Visit i on the displayed visit date.

Evidence snippet: "...symptom improved following treatment adjustment ..."

Visit information: Visit i; timestamp or relative visit index shown.

Surrounding note context: Relevant excerpt from the clinical-note timeline.

Judgments: Evidence match [True/False]; Question validity [True/False]; Answer correctness [True/False]; Evidence grounding [True/False]; Temporal consistency [True/False].

Figure 6: Schematic representation of the human evaluation interface. The example is illustrative and does not contain real patient text.

Annotation interface. The annotation interface presented one QA item at a time, including the question, answer, evidence snippet, visit information, and surrounding context, followed by five binary judgment fields corresponding to the criteria above. No patient identifiers or document identifiers intended for release were displayed in the final recorded annotation output. A schematic representation of the annotation interface is shown in Figure 6.

Human evaluator recruitment. The human quality evaluation was conducted by one clinical-domain researcher from the authors' institutional research network. We did not recruit crowdworkers, students, or participants through a public crowdsourcing platform. The evaluator reviewed a stratified sample of 1,000 QA pairs using the evaluation criteria described in Appendix C.1. No separate payment was provided for this evaluation, as it was conducted as part of the evaluator's normal research activities.

C.2 Quality Evaluation Stratified Data Distribution

Table 8 reports the distribution of the 1,000 QA-pair evaluated sample by reasoning type, difficulty, and visit group. The sample covers all temporal reasoning categories and is intentionally skewed toward long-history patients to stress-test quality under challenging longitudinal settings.

Reasoning type	Count	Difficulty	Count	Visit group	Count
Comparison across visits	256	Easy	196	2–10 visits	19
Progression	185	Medium	605	11–100 visits	116
Before–after	171	Hard	199	101–1000 visits	865
Treatment response	121				
Trend	113				
First/last occurrence	87				
Recurrence count	58				
Other	9				

Table 8: Distribution of the 1,000 QA-pair evaluation sample by reasoning type, difficulty, and visit group. The sample is intentionally skewed toward long-history patients to stress-test quality under challenging longitudinal settings.

Visit group	QA pairs
2–10 visits	200
11–100 visits	1,244
101–1000 visits	9,298

Table 9: QA-pair distribution by visit group.

Difficulty	**Count**
Easy	2,105
Medium	6,502
Hard	2,135

Table 10: Distribution of QA pairs by difficulty level.

D Additional Dataset Statistics

Table 9 summarizes the distribution of QA pairs across patient visit-frequency groups and shows that the dataset contains substantial coverage of medium- and long-history patient timelines.

Table 10 summarizes the distribution of QA pairs across difficulty levels, showing that the dataset is dominated by medium-difficulty questions.

Reasoning type	**Count**
Comparison	2,736
Progression	1,992
Before–after	1,837
Treatment response	1,299
Trend	1,218
First/last occurrence	937
Recurrence count	623
Other	100

Table 11: Distribution of QA pairs by temporal reasoning type.

Table 11 reports the distribution of temporal reasoning types, with comparison, progression, and before–after reasoning being the most common categories.

E Additional Experimental Setup

E.1 Retrieval Models

We evaluate four retrieval paradigms:

- **BM25**: sparse lexical retrieval baseline.
- **MedCPT**: clinical dense retrieval model based on contrastive pretraining over biomedical text.
- **NV-Embed-v2**: large general-purpose dense embedding model.
- **Hybrid BM25+MedCPT**: reciprocal rank fusion (RRF) combining sparse lexical retrieval and clinical dense retrieval with fusion parameter $k = 60$.

Retrieval is evaluated using Recall@K and Exact Coverage@K for $K \in \{5, 10, 20\}$. Recall@K measures whether at least one gold evidence visit is retrieved, while Exact Coverage@K requires retrieval of all gold evidence visits associated with a QA pair.

E.2 Reader Models

We evaluate three QA reader architectures:

- **ClinicalBERT**: short-context extractive QA model.
- **Longformer**: long-context extractive QA model.
- **Qwen2.5-7B-Instruct**: instruction-tuned generative language model.

In the end-to-end setting, readers receive the top retrieved visits concatenated in chronological order. In the oracle-evidence setting, retrieval is bypassed and readers receive the gold evidence visits provided in the dataset.

Component	Model / Method	Approx. Parameters	Use in Experiments
Sparse retriever	BM25	–	Lexical visit-level retrieval
Dense retriever	MedCPT	∼110M	Clinical dense retrieval
Dense retriever	NV-Embed-v2	∼7B	General-purpose dense retrieval
Hybrid retriever	BM25 + MedCPT	– / ∼110M	Reciprocal-rank fusion retrieval
Extractive reader	ClinicalBERT	∼110M	Extractive QA reader
Extractive reader	Longformer	∼149M	Long-context extractive QA reader
Generative reader	Qwen2.5-7B-Instruct	∼7B	Generative temporal answer synthesis
QA generation model	GPT-5.4-nano	Not publicly disclosed	Schema-constrained QA generation via API
LLM judge	GPT-5.5 Thinking	Not publicly disclosed	Supplementary LLM-based quality evaluation

Table 12: Models and methods used in dataset construction, retrieval, QA evaluation, and quality assessment. Parameter counts are approximate and correspond to commonly reported sizes for the referenced model families where available. Parameter counts for proprietary API models are not publicly disclosed.

E.3 Model Size and Computational Budget

Table 12 summarizes the model sizes, computing infrastructure, and approximate computational budget used in our experiments. All experiments were conducted on NVIDIA A100 GPUs. We evaluated pretrained models without additional task-specific fine-tuning, except for inference-time retrieval and question answering over the constructed TemporalEHR-QA benchmark.

For computational infrastructure, all local retrieval and reader experiments were run on NVIDIA A100 GPUs. The retrieval experiments treat each visit-level clinical note as an independent document within the corresponding patient's timeline. These steps were not run on local infrastructure, so we do not report them in GPU-hours.

F Additional Baselines Results

F.1 Additional Retrieval Baselines Results

Visit Group	BM25	MedCPT	NV-Embed
2–10 visits	99.5	99.5	99.5
11–100 visits	60.6	63.3	61.5
101–1000 visits	7.7	8.8	7.6

Table 13: Retrieval Recall@5 stratified by patient timeline length.

Table 13 reports retrieval Recall@5 stratified by patient timeline length. Retrieval performance is nearly perfect for short timelines with 2–10 visits, but drops sharply for patients with more than 100 visits, highlighting the difficulty of identifying relevant evidence in long longitudinal clinical histories.

Table 14 summarizes retrieval Recall@5 across difficulty levels. Retrieval performance decreases consistently from easy to hard questions, suggesting that the proposed difficulty labels align with increasing evidence localization complexity.

Difficulty	BM25	MedCPT	NV-Embed
Easy	22.0	23.0	22.0
Medium	14.2	15.1	14.2
Hard	12.5	13.1	12.3

Table 14: Retrieval Recall@5 stratified by difficulty level.

Reasoning Type	BM25	MedCPT	NV-Embed
First/last occurrence	20.7	21.5	20.2
Comparison across visits	18.2	19.3	18.2
Treatment response	14.8	15.6	14.6
Before–after	13.8	14.8	13.4
Recurrence count	13.6	14.3	13.3
Progression	13.7	14.8	13.8
Trend	11.2	12.4	11.5

Table 15: Retrieval Recall@5 stratified by temporal reasoning type.

Table 15 reports retrieval Recall@5 stratified by temporal reasoning type. Retrieval performance varies across reasoning categories, with first/last occurrence and comparison-across-visits being relatively easier, while progression and trend reasoning remain particularly challenging due to their reliance on temporally distributed evidence.

F.2 Additional QA Baselines Results

Difficulty	Token F1	Precision	Recall	BERT-F1	N
Easy	24.16	0.7465	0.7234	0.7341	2105
Medium	19.37	0.7319	0.6988	0.7144	6502
Hard	15.95	0.7279	0.6826	0.7042	2135

Table 16: QA performance stratified by difficulty level using the strongest overall pipeline (NV-Embed + Qwen2.5-7B).

Table 16 shows that performance degrades consistently as question difficulty increases. Both lexical and semantic metrics decline from easy to hard

Visit Group	Token F1	Precision	Recall	BERT-F1	N
2–10 visits	36.15	0.7955	0.7784	0.7860	200
11–100 visits	24.45	0.7514	0.7274	0.7385	1244
101–1000 visits	18.63	0.7303	0.6951	0.7118	9298

Table 17: QA performance stratified by patient visit-frequency group using the strongest overall pipeline (NV-Embed + Qwen2.5-7B).

questions, indicating that the proposed difficulty labels align well with actual reasoning complexity.

Table 17 shows that QA performance deteriorates substantially as patient timelines become longer. The sharp decline from short timelines to histories containing hundreds of visits highlights the difficulty of long-horizon temporal retrieval and reasoning over longitudinal clinical records.

Table 18 shows that performance varies across temporal reasoning categories. Questions involving first/last occurrence and treatment response are relatively easier, while progression, trend, and recurrence reasoning remain more challenging, likely because they require synthesizing evidence distributed across multiple visits and longer temporal spans.

Reasoning Type	Token F1	Precision	Recall	BERT-F1	N
Other	26.10	0.7739	0.7532	0.7626	100
First/last occurrence	23.98	0.7496	0.7235	0.7357	937
Treatment response	20.79	0.7412	0.7105	0.7250	1299
Before–after	20.30	0.7357	0.7055	0.7199	1837
Comparison across visits	18.42	0.7310	0.6957	0.7124	2736
Trend	19.97	0.7313	0.6908	0.7097	1218
Progression	17.85	0.7275	0.6927	0.7092	1992
Recurrence count	17.07	0.7231	0.6854	0.7031	623

Table 18: QA performance stratified by temporal reasoning type using the strongest overall pipeline (NV-Embed + Qwen2.5-7B).