Investigating LLM Capabilities on Long Context Comprehension for Clinical Question Answering

Feras AlMannaa[1*], **Talia Tseriotou**[2*], **Jenny Chim**[2], **Maria Liakata**[2,3]
[1]Istanbul Aydın University
[2]Queen Mary University of London
[3]The Alan Turing Institute
{firas.almanna3}@gmail.com,{t.tseriotou,m.liakata}@qmul.ac.uk

Abstract

Despite the impressive reported performance of Large Language Models (LLMs) on medical Question-Answering (QA), evaluation still overrelies on multiple-choice and extractive settings primarily assessing textbook knowledge, with limited work on open-ended QA over patient records. We present a comprehensive study investigating LLM long-context (LC) comprehension on longitudinal, patient-centric Electronic Health Records (EHRs) across multiple datasets, models and metrics, covering extractive, multiple-choice, and open-ended generative QA under different context formulation and Retrieval Augmented Generation (RAG) settings. We show that medically fine-tuned models offer no measurable benefit over base counterparts, smaller models exhibit stronger memorization, and longitudinal supplementary notes aid generative QA. Reasoning models gain on complex multi-note QA under LLM Correctness but not stricter reference-aligned metrics, while hybrid RAG with reranking outperforms Full-Context (FC) prompting. A question-category analysis and multi-faceted metric evaluation further reveal where RAG and FC excel and struggle, with RAG falling behind in longitudinal reasoning, thus uncovering systematic failures pointing to the need for reasoning-appropriate metrics.

1 Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Grattafiori et al., 2024; Yang et al., 2024) perform impressively on medical tasks (Borgeaud et al., 2022; Nori et al., 2023), achieving superhuman scores on United States Medical Licensing Examination (USMLE)-style exam question-answering (QA) (Chen et al., 2023b; Tang et al., 2024; Pal and Sankarasubbu, 2024; Singhal et al., 2025). Yet evaluation is mostly based on multiple-choice QA (MCQA) (Liévin et al., 2024; Xiong et al., 2024; Chen et al., 2025b; Singhal et al., 2025) which doesn't reflect performance in complex tasks (Arias-Duart et al., 2025), such as open-ended medical QA (Sandeep Nachane et al., 2024). Furthermore, medical board exams designed to assess professional knowledge and decision-making rely primarily on textbook knowledge which is likely available during pre-training (Chen et al., 2025a). With clinicians facing increasing workloads (Martinez et al., 2024), LLM-based QA on patient EHRs offers promise, yet existing deployment studies rely on shallow, category-triggered context and narrow clinical evaluation (Garcia et al., 2024; Bootsma-Robroeks et al., 2025). Systematic assessment on expert-curated EHRs would signal how well models handle real-world clinical complexity, yet research focuses largely on dataset creation (Yang et al., 2022) or specialized models (Fleming et al., 2024) rather than on assessing LLMs on realistic long-context clinical settings. Related work has examined LLMs for diagnostic prediction over EHRs (Pan et al., 2025; Albassam et al., 2025), but is limited to specific diagnoses or single datasets and to short-form accuracy-based evaluation.

Most EHR-based benchmarks consider single-document contexts (Pampari et al., 2018; Yue et al., 2021), yet in practice relevant information may span multiple documents. Reasoning over multiple long, temporally-dependent documents raises questions around Long-Context (LC) limitations and evidence positioning sensitivity (Liu et al., 2024; Xiao et al., 2023; Adams et al., 2025). Moreover existing LC medical QA work is either synthetic, MCQA-based and short (Adams et al., 2025) or small and not human-validated (Fan et al., 2025a).

We address this gap by investigating LLM capabilities on longitudinal, patient-centric EHRs with human-validated QA pairs. While recent LLMs push the frontiers of context size, performance does

* Indicates equal contribution.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1188–1213
November 6-10, 2026 ©2026 Association for Computational Linguistics

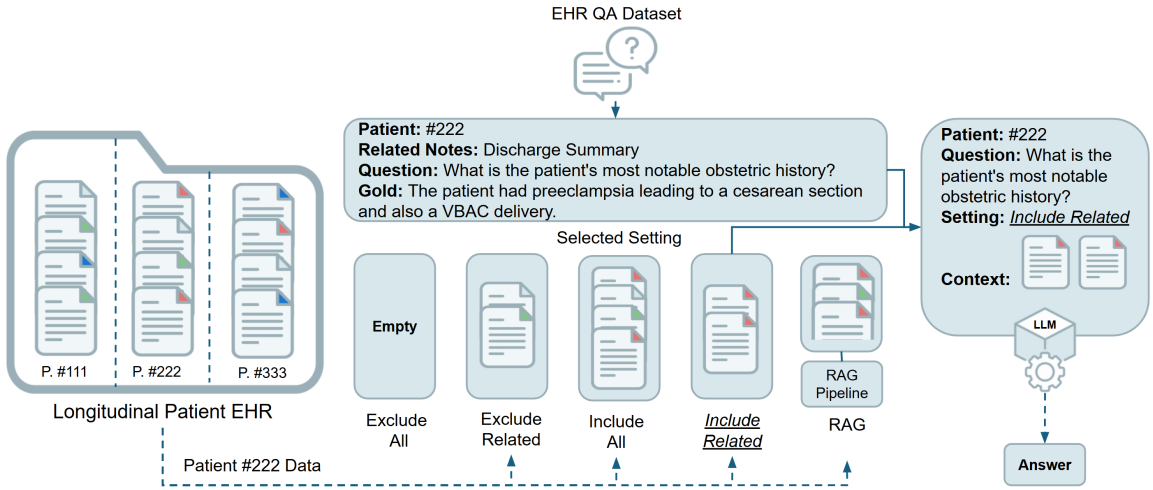


Figure 1: Overview of our Approach. Given a patient question grounded in EHRs, the most related note (per original dataset) and an expert-verified gold answer (used only for evaluation), we augment the relevant note with the patient's full longitudinal EHR history from MIMIC (§3.1). We evaluate four prompting settings: (a) *Include Related*: only the most related note(s), (b) *Include All*: all longitudinal notes, (c) *Exclude Related*: supplementary notes only, (d) *Exclude All*: no notes. We also examine two temporally ordered RAG strategies over all notes: (i) *RAG*: retrieved 512-token chunks, (ii) *RAG HIR*: hierarchical parent note inclusion by substituting retrieved chunks with their full parent note (§3.4, 4.2). Evaluation spans two extractive and two open-ended generative QA datasets.

not scale with increased context (Levy et al., 2024; Modarressi et al., 2025). RAG (Lewis et al., 2020) offers a cost-efficient alternative by selecting relevant context on demand, with reported gains over Full-Context (FC) across various settings (Zhang et al., 2024a; Xu et al., 2024a; Bai et al., 2025), though its advantage peaks at 32K for large LLMs (Leng et al., 2024). Models still struggle with specialized content such as EHRs (Fan et al., 2025b; Myers et al., 2025; AlMannaa et al., 2026), leaving Full-Context (FC) vs RAG superiority in LC clinical EHR-based QA an open question. Prior works comparing RAG and FC (Li et al., 2024b) or examining RAG's limits (Leng et al., 2024) relied on a single hand-picked dense retriever, which may underestimate RAG's potential. EHR-specific retrieval work has shown the complementarity of dense and sparse retrievers in medical tasks (Fan et al., 2025b; Myers et al., 2025) with hybrid retrieval outperforming both (Stuhlmann et al., 2025). We design a hybrid RAG pipeline composed of well-established robust components as a principled baseline enabling a fair comparison against FC prompting in LC settings. We make the following contributions:

- We assess how LLMs of different capabilities and sizes perform on Long-Context clinical QA given longitudinal EHR patient notes, covering three task formulations: 1) MCQA, 2) Extractive, and 3) Open-ended generative QA (§4.1, §4.2) across a variety of LC and RAG settings (Fig. 1).
- We perform comprehensive experiments evaluating different note inclusion strategies, examining their effect on performance and model memorization (§3.1).
- We evaluate RAG against Full-Context (FC) prompting over longitudinal EHR patient histories across task formulations and context sizes, demonstrating RAG superiority across tasks (§5.1, §5.2), through a designed hybrid RAG pipeline motivated by biomedical retrieval literature (§3.3), yielding insights into long-context EHR comprehension.
- Our multi-faceted evaluation methodology covers both single- and multi-document long-form reasoning settings, while a category-level quantitative analysis reveals where RAG and FC respectively excel and struggle, with RAG falling behind in longitudinal reasoning (§5.1, §5.2, §5.3).

2 Related Work

Leveraging Long Context: While increasing long-context window capabilities of LLMs, e.g. GPT-4o (Hurst et al., 2024) (128K), Claude 3 (Anthropic, 2023) (200K), Gemini 1.5 (Team et al., 2024a) (1 million), boost their Long-Context performance, they still struggle in general purpose open-form QA beyond short passages (Zhang et al., 2024b; Chen et al., 2025c). Despite positional embedding

techniques like RoPE (Su et al., 2024), YaRN (Peng et al., 2023) and Position Interpolation (Chen et al., 2023a), QA task performance remains low even for 32K contexts (Chen et al., 2025c), with LLMs particularly challenged by long-range dependencies requiring reasoning (Li et al., 2024a; Dong et al., 2024; Bai et al., 2024; Liu et al., 2024; Hsieh et al., 2024). The performance gap between open-source and closed-source models (Li et al., 2024a) further raises concerns for patient-centric settings where privacy matters. Fan et al. (2025a) show notable performance drops in medical QA up to 200K tokens, yet despite the prominence of longitudinal context in real-world patient data, little work investigates long-form LLM capabilities.

Retrieval-augmented Generation (RAG) has been leveraged in a prompt-based plug-in manner to address black-box Long-Context challenges (Yu et al., 2023). LLM-based retrievers like BGE M3 (Chen et al., 2024) and self-reflective approaches like Self-RAG (Asai et al., 2024) have shown notable gains in open-domain QA. In the biomedical domain, MedCPT (Jin et al., 2023) and BMRETRIEVER (Xu et al., 2024b) achieve strong retrieval performance, while Med-RAG (Zhao et al., 2025) and Self-BioRAG (Jeong et al., 2024) further improve EHR-based diagnostics through knowledge graph reasoning and self-reflection respectively. Yet despite strong in-domain retrieval capabilities, models still struggle with specialized medical content such as EHRs (Fan et al., 2025b; Myers et al., 2025), motivating a thorough examination of RAG versus FC over realistic long-context patient-oriented QA. Instead of relying on a single hand-picked retriever, we design a literature-grounded principled hybrid RAG pipeline (Ryan et al., 2026; Stuhlmann et al., 2025; Sohn et al., 2025) to enable a fair and robust comparison against FC on longitudinal EHR-based QA (see §3.3).

3 Approach

Since strong medical QA performance offers the potential of LLM integration in clinical workflows, we aim to uncover LLM limitations given complex patient-oriented scenarios. We focus on **content-comprehension** settings using longitudinal EHRs, thus emphasizing relevant patient-centric **long-context** across time and documents.

Our comprehensive approach spans across expert annotated datasets covering **three task formulations**: 1) MCQA, 2) Extractive, and 3) Open-ended QA. We assess each dataset across **four settings with varying note relevance** (§3.1).

3.1 Context Inclusion Formulation

Most datasets use a single note/report or up to a few discharge summaries, leaving a large amount of *longitudinal patient supplementary* content unused in the EHR history. We aim to assess the significance of the patient's longitudinal Full-Context (FC), evaluating a realistic clinical setting with no a-priori knowledge of the most relevant note(s). Therefore we leverage the MIMIC resources to augment the context with each patient's longitudinal notes[1], resulting in Long Context (LC) from the entire patient's note history (*Include All*). We then investigate how LLMs leverage long-form context to answer patient-specific questions.

We explore four data inclusion scenarios depicted in Fig. 1, allowing comprehensive analysis:

1. **Exclude All**: Exclusion of all notes - assessing model note memorization.
2. **Exclude Related**: Exclusion of the identified most relevant notes - assessing the informational value of only the supplementary patient notes for QA, without including the note(s) identified as most relevant, thus isolating the contribution of longitudinal supplementary context.
3. **Include All**: Inclusion of all patient notes - assessing LC processing of varying importance.
4. **Include Related**: Inclusion of only the most relevant note(s) (DS, CN or RR) for the question from the patient's MIMIC context as defined by each dataset's release (see *Answer Location* in Table 1) - assessing the model's ability to leverage the most relevant context.

These settings enable direct comparison between memorization, supplementary context utility and full versus targeted context inclusion, spanning from no context to the full longitudinal history.

3.2 Dataset Selection Criteria

The following criteria were used for benchmark selection (full comparison in Table 12):

Beyond MCQA: Medical MCQA benchmarks i.e. PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022) evaluate option selection instead of grounded synthesis and risk pretraining leakage from textbooks, which can inflate scores, providing little signal towards noisy

[1]MIMIC preprocessing is described in Appendix C.2.

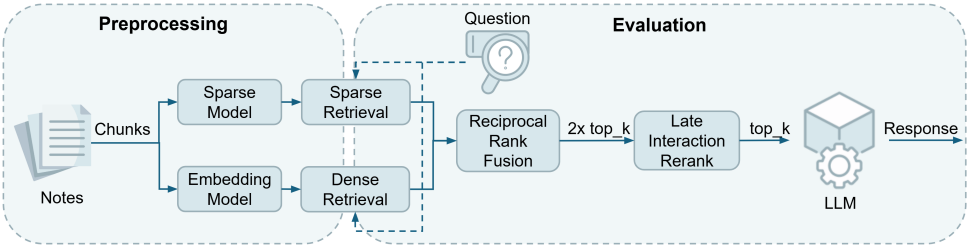


Figure 2: RAG Pipeline: clinical notes are split into 512-token chunks, encoded using dense (Qwen3-Embedding-8B) and sparse (BM25) embeddings, and indexed in Qdrant. At evaluation time the question retrieves 2×top-k chunks from each retriever, Reciprocal Rank Fusion (RRF) combines them into a total of 2×top-k chunks, and Late Interaction Reranking yields the final top-k chunks, sorted chronologically before being passed to the LLM.

clinical settings. We go beyond MCQA evaluating on both extractive and open-ended QA.

Long-Context Documents: To address LLM capabilities in handling real-world LC we seek benchmarks extending well-beyond 8K, with inclusion of patient-specific content.

Expert Annotated QA pairs grounded on multi-document EHRs were our choice, to avoid clinical reliability issues like non-expert curation (Fan et al., 2025a) or synthetic data (Adams et al., 2025).

Patient-centric Longitudinal Content: Prioritizing clinically-relevant QA with reasoning, we focus on MIMIC-based EHR datasets (Johnson et al., 2016, 2023), offering timestamped notes across document types with consistent format.

3.3 RAG Approach

Retrievers are distinguished into *sparse* and *dense*. Sparse retrievers such as BM25 (Robertson and Zaragoza, 2009) and Splade (Formal et al., 2021) use overlapping terms to match queries with snippets, while dense retrievers perform matches via semantic similarity in the embedding space. Using LLM-based embeddings for dense retrieval has become standard practice.

Research on biomedical MCQA (Wang et al., 2024) and biomedical document retrieval (Luo et al., 2022) has shown that hybrid approaches combining sparse and dense retrieval outperform single component counterparts. Hybrid sparse-dense retrieval with Reciprocal Rank Fusion (RRF) has further demonstrated strong performance across domains (Arivazhagan et al., 2023; Lee et al., 2023), while recent work across biomedical QA shows that hybrid retrieval outperforms both sparse and dense retrievers alone (Stuhlmann et al., 2025), with hybrid retrieval with RRF specifically demonstrating superiority on PubMedQA and MedQA (Ryan et al., 2026).

Additionally, related EHR work has further shown that no single embedding model dominates across medical retrieval tasks (Myers et al., 2025), with dense and sparse retrievers offering complementary benefits depending on the setting and dataset (Fan et al., 2025b). Finally, re-ranking of retrieved snippets also exhibits superior performance in biomedical applications (Jin et al., 2023; Wang et al., 2024; Sohn et al., 2025). FC versus RAG comparisons (Li et al., 2024b; Leng et al., 2024) rely on a single dense retriever, potentially underestimating RAG's potential. Provided substantial evidence for the superiority of hybrid retrieval over single-component counterparts in this domain, we leverage a hybrid approach followed by re-ranking, presented in Figure 2, focusing on our primary goal on evaluating LLM capabilities under different context settings on long longitudinal EHR patient data. Clinical notes are segmented into 512-token chunks, forming the document collections $D = \{d_1, d_2, \ldots, d_n\}$. The question respectively forms the query Q.

Sparse retrieval: we employ a lexical retriever as a ranking function of each chunk d_j [2]:

$$S_{\text{sparse}}(Q, d_j) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f_{q_i,d_j}\,(k_1+1)}{f_{q_i,d_j} + k_1\left(1-b+b\cdot\frac{|d_j|}{\text{avgdl}}\right)} \tag{1}$$

Dense retrieval: we use the same encoder, E, for the query and the chunks and capture semantic similarity of such pairs using the cosine similarity:

$$S_{\text{dense}}(Q, d_j) = \frac{E(Q) \cdot E(d_j)}{\|E(Q)\|\,\|E(d_j)\|} \tag{2}$$

Reciprocal Rank Fusion (RRF): combines the top-k chunks from each retriever, through $\text{rank}_s(d_j)$ denoting the rank assigned to document d_j by re-

[2] Here terms document and chunk are used interchangeably.

triever $s \in \{\text{dense}, \text{sparse}\}$, using a smoothing constant k_{RRF}:

$$\text{RRF}(d_j) = \sum_{s \in \mathcal{S}} \frac{1}{k_{RRF} + \text{rank}_s(d_j)} \quad (3)$$

Late Interaction Reranking: measures how well every Q-token, $\mathbf{q}_i$, is supported by at least one document token, $\mathbf{d}_{j,r}$, to produce the top-k set:

$$S_{\text{MaxSim}}(Q, d_j) = \sum_{i=1}^{|Q|} \max_{1 \le r \le |d_j|} \langle \mathbf{q}_i, \mathbf{d}_{j,r} \rangle \quad (4)$$

Document Ordering: we sort retrieved chunks chronologically by note timestamp rather than retrieval score, as temporal ordering yields superior long-context performance (Yu et al., 2024).

3.4 Evaluation Methodology

Evaluating medical QA systems has traditionally relied on single metrics and MCQA benchmarks (Jin et al., 2019, 2021; Pal et al., 2022), which probe memorized knowledge and underrepresent the complexity of clinical reasoning, and EHR-grounded context. Aiming for a clinically meaningful assessment, we adopt a multi-dimensional, reference-based scheme that complements lexical overlap with embedding-based semantic similarity and domain-adapted Natural Language Inference (NLI). In line with emerging clinical evaluation practices (MEDIC (Kanithi et al., 2024)) and frameworks (ArchEHR (Soni et al., 2025)) and the rise of LLM-as-a-judge, we employ a calibrated rubric and we prioritize widely available metrics while cross-checking signals for robustness.

Our datasets comprise gold-standard QA pairs that are either expert-verified or expert-generated, thus enabling a clinically centric evaluation:

METEOR (Banerjee and Lavie, 2005) to capture *surface-level* and *lexical similarity*.

BERTScore (Zhang et al., 2019), computed with Clinical BioBERT embeddings (Alsentzer et al., 2019) to assess *semantic similarity*.

NLI Scores, to capture the logical relationship between reference and candidate answers. We use a domain-adapted NLI model (Deka et al., 2023) and then measure *Factual Consistency* using the probability of non-contradiction (Song et al., 2024) and *Factual Precision* using the entailment probability.

LLM-as-a-judge, employed with a 5-point scale rubric on three aspects: *Correctness* capturing factual consistency with respect to the gold answer, penalizing contradictions while allowing compatible additional information; *Completeness* capturing recall by assessing the predicted answer's coverage of information present in the reference; *Faithfulness* capturing precision assessing if the predicted answer only contains information that is supported by the reference. Prompts are provided in Appendix I.

Accuracy reported on the MCQA task.

Our metric selection is grounded in correlation analysis (Figures 10, 11 in Appendix F.1), prioritizing complementary metrics with lower inter-correlation to maximize signal diversity and minimize redundancy. Metrics correlate stronger within families, while METEOR and LLM Faithfulness are also highly correlated with BERTScore. In the long-context high-correctness regime, correlations are weaker. Thus, in the main tables we report one metric per family for a comprehensive yet non-redundant view across complementary dimensions.

4 Experiments

4.1 Datasets

We leverage MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023), comprising de-identified EHRs covering hospital admissions and ICU stays, as the underlying context. They support single- and multi-note settings and include: clinical notes (CN), discharge summaries (DS) and radiology reports (RR) per patient over time.

Our four selected QA datasets only use QA pairs curated or validated by clinical experts to ensure medical validity. We provide a brief dataset overview below, also summarized in Table 1:

CliniQG4QA (Yue et al., 2021): An extractive span benchmark on MIMIC-III comprising 1,287 QA pairs generated or verified by three clinical experts. We only use the expert-annotated set.

RadQA (Soni et al., 2022): An extractive dataset of 3,074 physician-crafted questions and 6,148 answer spans from Findings and Impressions sections of radiology reports in MIMIC-III, annotated by two human annotators. After filtering unanswerable pairs we retain 3,509 pairs.

EHR-DS-QA (Kotschenreuther, 2024): A synthetic QA corpus generated by two LLMs on discharge summaries and clinical notes from MIMIC-IV. Of 506 physician-verified pairs, we retain 478 marked as correct.

EHRNoteQA (Kweon et al., 2024): Built on MIMIC-IV, comprising 962 QA pairs authored by GPT-4 and iteratively refined by three clinicians, each corresponding to multiple discharge

Dataset	Context Data	QA Pairs	Patients	Mean/Max Context	Task Types	Answer Location	Reasoning	Question Generation	Answer Annotation
CliniQG4QA	MIMIC-III	1,287	36	4K / 7K	Extractive	Clinical Note	×	Experts	Clinical experts (3)
RadQA		3,509	80	6K / 14K		Radiology Report	Single-Note	Human-generated	Physicians (2)
EHR-DS-QA	MIMIC-IV	478	70	46K / 131K	Open-ended	Clinical Note/Dis. Summary	×	LLMs	Physician-verified (1)
EHRNoteQA		962	962	9K / 39K	MC,Open-ended	Dis. Summaries	Multi-Note	LLM	Clinician-refined (3)

Table 1: EHR QA Dataset Statistics. Mean/Max Context corresponds to the *Include All* setting.

summaries (∼2.3 per patient) and supporting both open-ended and MCQA formats.

Context Bins: For *Include All* we report performance across four token bins: 0–8K, 8–16K, 16–32K, and 32–128K. As shown in Appendix C.3, both *Include Related* and *Exclude Related* yield shorter contexts across all bins, most notably in the 16–32K bin for *Include Related*.

4.2 Models and Experimental Setup

LLMs: To explore the effect of model size, domain specialization and reasoning capabilities we study four open Qwen2.5 (Team et al., 2024b) single-family models, ensuring a controlled comparison:

- *Qwen2.5-7B-Instruct*
- *HuatuoGPT-o1-Qwen-7B* (Chen et al., 2025b): medical LLM based on Qwen2.5-7B-Instruct
- *Qwen2.5-32B-Instruct-128K*
- *QwQ:32B*: reasoning model

Evaluation: Through comparisons against Selene-8B (Alexandru et al., 2025) and Prometheus-8x7B v2.0 (Kim et al., 2024), we finally selected QWEN-2.5-32B-INSTRUCT for the LLM-as-a-judge for stability and agreement (see Appendix D.3).

Retrieval Settings: Based on biomedical QA retrieval research that demonstrates the superiority of combining sparse and dense retrievers (see §3.3), we designed a hybrid retrieval approach to handle the semantic, lexical, and query complexity. The selected model components are listed below:

- *Dense Retrieval* using the open embedding model Qwen3-Embedding-8B (Zhang et al., 2025).
- *Sparse Retrieval* using BM25, selected after ablating two sparse retrievers (see Table 11).
- *Late Interaction Reranking* with Reason-ModernColBERT, a ColBERT (Khattab and Zaharia, 2020) model finetuned on the ReasonIR dataset (Shao et al., 2025) demonstrating high performance on a reasoning-intensive retrieval benchmark (Su et al., 2025).

We explore two chunk inclusion strategies: i) direct chunk inclusion (*RAG*), ii) hierarchical whole parent note inclusion of the retrieved chunk for full-note context view (*RAG HIR*). Under both settings, the retrieved content is sorted chronologically using the corresponding MIMIC note creation timestamps before being concatenated into the prompt (§3.3). Therefore the model receives a patient timeline. Retrieval is based on 512-token chunks selected through chunking strategy experimentation (Appendix D.2, Table 14).

Experimental Setup: LLM inference and prompt formulation specifics are in Appendices A and C.4.

5 Results and Discussion

5.1 Main Results

Performance for the different inclusion settings across datasets and models is shown in Table 2. Statistical significance of the key claims drawn from these results is examined through Philipp Koehn's paired bootstrap resampling tests (Appendix E).

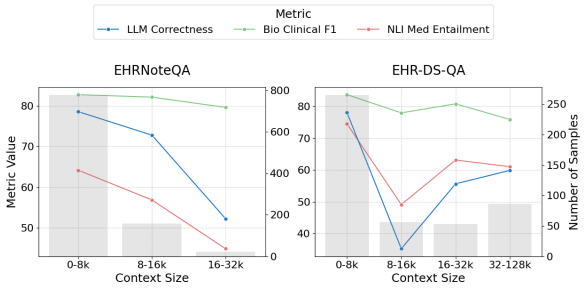


Figure 3: Metric performance over context size on *Include All* for Qwen2.5:32B-Instruct.

Model Size, Tasks and Inclusion Settings: Consistent with general findings, larger models perform better across all tasks (Table 2), with the performance gap particularly pronounced in tasks requiring reasoning or information synthesis from multiple sources, namely EHRNoteQA and RadQA. However, smaller models perform better in settings excluding relevant information or completely removing the context, suggesting stronger memorization as also strongly suggested by Philipp Koehn's paired bootstrap resampling significance test (Koehn, 2004) reported in Table 16 (performed on open-ended generative datasets). Unlike the

Model	Setting	EHRNoteQA				EHR-DS-QA			RadQA			CliniQG4QA		
		Open-ended			MC	Open-ended			Extractive			Extractive		
		LLM	NLI	F1	Acc.	LLM	NLI	F1	LLM	NLI	F1	LLM	NLI	F1
HuatuoGPT-o1 7B	Exclude All	7.85	**23.36**	**68.82**	51.08	**26.65**	**33.86**	72.21	25.06	27.90	63.78	15.96	23.01	64.44
Qwen2.5 7B		**9.73**	18.18	68.41	51.46	24.79	33.00	72.35	**39.08**	**29.24**	**70.73**	**27.02**	**26.30**	**74.06**
Qwen2.5 32B		0.71	11.30	66.56	58.34	20.16	20.23	**72.67**	3.86	2.88	10.35	0.11	0.19	2.16
QwQ 32B		2.68	12.23	66.98	**59.24**	24.16	23.80	71.70	1.51	6.44	15.29	0.93	5.62	14.77
HuatuoGPT-o1 7B	Exclude Related	26.62	**40.26**	72.57	58.35	28.49	**33.98**	71.84	24.39	28.28	63.33	15.70	24.13	64.15
Qwen2.5 7B		25.24	28.64	72.72	59.86	**29.90**	28.17	73.35	**38.86**	**29.07**	**70.72**	**27.01**	**25.95**	**74.09**
Qwen2.5 32B		25.59	23.67	73.01	58.72	22.16	22.35	73.15	3.77	2.78	10.02	0.13	0.20	2.17
QwQ 32B		**27.46**	23.01	**74.35**	**60.15**	27.19	23.98	**73.51**	1.14	7.11	17.48	0.99	4.84	15.75
HuatuoGPT-o1 7B	Include All	72.60	45.45	78.21	78.94	59.76	**63.66**	77.14	63.92	49.41	76.38	67.85	52.45	80.08
Qwen2.5 7B		70.49	38.11	79.95	78.81	**62.49**	61.00	79.14	62.56	50.13	76.65	68.89	54.13	81.27
Qwen2.5 32B		67.88	**55.33**	**81.50**	**90.97**	57.21	61.96	**79.59**	**67.74**	50.10	**77.55**	**80.49**	59.22	83.80
QwQ 32B		**75.52**	47.14	81.02	89.25	57.68	60.54	78.90	67.46	**53.26**	77.47	78.95	**66.00**	**83.94**
HuatuoGPT-o1 7B	Include Related	70.73	39.30	79.71	76.87	63.46	60.65	77.20	64.37	49.19	76.48	70.47	53.19	79.83
Qwen2.5 7B		74.44	39.15	80.78	80.9	**64.76**	**65.14**	80.28	63.01	50.25	76.71	69.06	54.87	81.29
Qwen2.5 32B		76.01	**61.41**	**82.48**	**90.33**	59.12	64.86	**80.70**	67.76	50.78	77.67	**79.81**	59.21	83.55
QwQ 32B		**82.03**	47.19	82.36	87.57	61.31	61.68	79.70	**67.79**	**55.48**	**77.69**	79.74	**65.86**	**84.35**

Table 2: Results for Full Context in each setting across datasets. **Bold** is best and underlined the second best model in each Setting and Metric. Orange highlights the global best across all Settings per Metric. *LLM* corresponds to LLM Correctness, *NLI* to NLI Entailment and *F1* to BERTScore (BioBERT) F1.

Model	Setting	EHRNoteQA				EHR-DS-QA		
		Open-ended			MC	Open-ended		
		LLM	NLI	F1	Acc.	LLM	NLI	F1
HuatuoGPT-o1 7B	Include All	70.68	39.52	77.26	75.36	50.25	59.01	75.30
	Include Related	66.88	29.62	79.37	72.38	56.51	54.71	75.29
	RAG 5	59.94	**61.34**	78.44	75.67	57.57	61.53	78.22
	RAG 10	61.03	48.51	77.99	73.4	58.08	62.45	**78.61**
	RAG 15	60.84	52.53	77.92	74.97	**61.95**	**64.84**	77.68
	RAG HIR 3	70.13	56.91	77.99	70.81	50.06	53.82	76.17
	RAG HIR 5	**70.92**	57.51	**79.52**	72.38	48.46	55.21	76.66
	RAG HIR 7	70.80	54.61	79.35	**80.46**	55.99	57.12	77.35
Qwen2.5 7B	Include All	66.86	28.36	79.23	75.04	54.26	52.78	77.07
	Include Related	72.37	29.50	80.34	78.82	58.60	60.79	78.85
	RAG 5	59.54	40.33	78.82	75.67	54.17	56.11	**79.82**
	RAG 10	64.03	45.02	79.05	73.40	57.11	**62.76**	79.60
	RAG 15	58.07	52.08	79.20	75.60	**59.25**	60.32	79.65
	RAG HIR 3	65.54	46.34	80.37	76.93	50.73	53.12	78.65
	RAG HIR 5	**72.68**	**54.69**	80.17	79.76	53.59	52.56	77.66
	RAG HIR 7	67.55	46.41	**80.69**	**84.31**	52.10	52.39	78.26
Qwen2.5 32B	Include All	62.54	50.90	80.87	89.79	50.27	57.78	78.22
	Include Related	**74.16**	**59.11**	82.17	88.78	51.88	59.78	79.48
	RAG 5	60.35	46.36	80.28	80.46	58.02	58.91	80.29
	RAG 10	67.00	54.54	80.47	77.87	**60.52**	60.31	80.24
	RAG 15	73.45	53.94	80.84	81.02	55.68	61.39	**80.71**
	RAG HIR 3	67.48	57.17	**82.41**	83.36	55.13	**63.32**	80.55
	RAG HIR 5	68.86	54.87	81.59	90.74	51.20	60.73	80.12
	RAG HIR 7	72.00	47.33	81.84	**91.68**	52.82	57.27	80.21

Table 3: Open-ended and MCQA Results for Context of 8K+ tokens including LC and RAG Methods. **Bold** is best and underlined the second best performance per Model and Metric. Orange highlights the global best across all models per Metric.

rest of the tasks, Multiple-Choice (MC) shows strong evidence of memorization with a performance of 59.24% in the *Exclude All* setting, while MC demonstrates 90%+ performance, open-ended and extractive tasks remain challenging for LLMs.

The *Exclude All* setting primarily assesses model memorization, with Qwen2.5-7B and HuatuoGPT-o1-7B outperforming Qwen2.5-32B further indicating stronger memorization in smaller models. The *Exclude Related* setting, which includes only supplementary patient history excluding the most relevant notes, yields near-identical performance to *Exclude All* across all models and sizes for the purely extractive RadQA and CliniQG4QA datasets, where the answer is contained within the single related note, rendering supplementary context uninformative. By contrast, for the open-ended generative datasets, *Exclude Related* outperforms *Exclude All* by a margin, demonstrating the usefulness of supplementary patient material as also strongly supported by the significance tests in Table 17. The *Include Related* setting outperforms *Include All* on most datasets and metrics except for MC, while the gap between the two suggests that LLMs struggle to identify the most relevant information in the full *Include All* context.

Reasoning and Fine-tuning: *The QwQ:32B general reasoning model* showed improved results on Table 2 over other models overall on RadQA, EHRNoteQA (the reasoning datasets), suggesting that reasoning models can offer an advantage in tasks that demand complex inference across multiple sources of information, particularly at larger model sizes where reasoning capabilities can be effectively leveraged. Significance testing (Table 18) confirms that QwQ-32B significantly outperforms Qwen2.5-32B on LLM Correctness for multi-note QA (EHRNoteQA) but not on NLI or BioClinical F1, nor on any metric for single-note QA (EHR-DS-QA). This suggests reasoning models generate

Model	Setting	RadQA			CliniQG4QA		
		LLM	NLI	F1	LLM	NLI	F1
HuatuoGPT-o1 7B	Include All	35.60	**48.81**	75.61	43.73	50.44	78.97
	RAG 5	**36.61**	29.51	**76.61**	**44.20**	**54.51**	**79.54**
Qwen2.5 7B	Include All	36.10	**48.59**	75.92	44.55	53.26	80.04
	RAG 5	**37.09**	28.39	**77.10**	**47.03**	**58.36**	**81.73**

Table 4: Extractive QA Results for Context of 4K+ tokens. **Bold** is best performing per Model and Metric. Orange highlights the global best per Metric.

answers that go beyond the gold reference while remaining factually compatible, penalized by the reference alignment of NLI and BioClinical F1 (§5.3).

By contrast, *the medically fine-tuned reasoning model, HuatuoGPT-o1 7B*, is statistically indistinguishable from the instruction-tuned Qwen2.5:7B-Instruct across metrics and open-ended datasets (Table 15). We thus report an *absence of measurable benefit* from medical fine-tuning, consistent with reported reduced performance of biomedical LLMs (Dorfner et al., 2024; Dada et al., 2025).

Context Size: Fig. 3 shows that performance decreases with increased context size on open-ended QA, particularly for multi-note reasoning (EHRNoteQA), with similar trends across models (see Appendix, Fig. 8). Notably, EHR-DS-QA exhibits a performance dip at 8-16K tokens across metrics and models, attributable to data noise and some ambiguous gold answers, consistent with observations by Fan et al. (2025a) and Wang et al. (2026).

RAG Performance: Table 3 shows two RAG settings: *RAG* x and *RAG HIR* x, with x corresponding to the number of child 512-token chunks retrieved[3]. RAG demonstrates clear improvement on the single-note EHR-DS-QA task, where direct chunk inclusion consistently performs best. *RAG HIR* is superior for the more complex multi-note reasoning of EHRNoteQA including MC for the smaller models, while larger models benefit less from retrieval in complex reasoning settings. Significance testing in Table 19 performed on the best on average RAG strategy per dataset and model confirms these findings. Impressively, RAG surpasses *Include Related* performance across most settings, suggesting that effective retrieval can match or exceed having the most relevant note available a priori, with the exception of the multi-note reasoning dataset and larger models. Findings are consistent across 8K+ context sizes (see Fig. 9), with smaller models benefiting most. The medically fine-tuned reasoning model HuatuoGPT-o1 7B scores the best results on multiple metrics across the open-ended datasets, showing that RAG could benefit even more the smaller models that struggle in Full-context (FC) settings *Include All*. In Table 4 we only evaluate 7B sized models using a single RAG setting, due to the task nature and limited context size of the extractive datasets. While the extractive datasets have shorter contexts overall, RAG still yields the best performance.

[3]We excluded QwQ:32B due to time and budget constraints

5.2 Quantitative Analysis: RAG vs FC

We first isolate challenging cases of metric disagreement (LLM Correctness, NLI Entailment, and BioClinical F1) across the open-ended QA EHRNoteQA and EHR-DS-QA. Metrics were first z-score standardized and samples were gathered for cases where some metrics belonged to the top 60th percentile while another metric fell within the bottom 40th percentile. The resulting 41 samples were manually categorized across QA types to determine which answer between RAG and FC is more reliable. We form the following rubric based on the manual analysis and show results in Table 5:

- **Specific Entity Extraction (SEE):** Requiring retrieval of high cardinality facts such as medications and lab values or surgical procedures.
- **Temporal & Longitudinal Reasoning (TLR):** Requiring tracking of changes over time, chronological event ordering or visit comparisons.
- **Clinical Nuance & Status (CNS):** Questions regarding discharge instructions, patient mental status or synthesized clinical course summaries.
- **Absence or Negative Information (ANI):** Requiring the identification of absent information.

Category	Total	Favored by RAG	Favored by FC
SEE	18	**12 (66.7%)**	6 (33.3%)
TLR	10	3 (30.0%)	**7 (70.0%)**
CNS	10	**6 (60.0%)**	4 (40.0%)
ANI	3	**2 (66.7%)**	1 (33.3%)
Total	41	**23 (56.1%)**	18 (43.9%)

Table 5: Manual categorization in open-ended QA of challenging metric disagreement cases showing where RAG vs FC is more reliable.

RAG shows higher performance across the *SEE* and *CNS* categories, while FC leads in the *TLR* one, demonstrating that while RAG effectively preserves the relevant information for specific facts and clinical nuance challenging cases, it falls be-

hind when temporal reasoning is required disrupting continuity. A further qualitative analysis providing insights on cases where RAG or FC perform better (see Appendix §H) showcases findings on par with the quantitative ones, with the exception of *ANI* where FC has more promising although inconsistent performance.

We further scale up the above rubric by prompting `Gemini-3-flash` with Chain-of-Thought reasoning to automatically categorize all 8K+ token instances in the multi-note reasoning EHRNoteQA dataset, computing FC versus RAG success rates via min-max normalization across LLM Correctness, NLI Entailment, and Bio F1. Human evaluation of 20 randomly sampled instances shows 85% categorization accuracy. In Table 6, we show that RAG has superior performance in the long-context categories of *CNS* and *ANI*, showing superior signals of consistency and factual grounding where clinical nuance, discharge status or absence of information must be identified. By contrast, FC has stronger performance in *TLR*, where tracking chronological events and longitudinal changes across the full context is more reliable over note fragmenting. FC and RAG are equally good in Specific Entity Extraction *SEE*. Retrieval is the better default over long patient histories, winning in *CNS* and *ANI* both manually and at scale, except for temporal and longitudinal reasoning question *TLR*, where FC prevails.

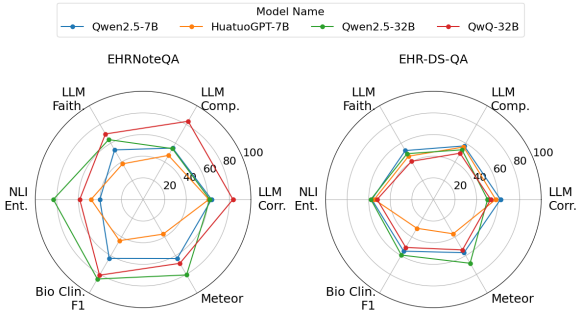


Figure 4: Model performance across metrics for open-ended QA, normalized with min-max per metric.

		Favored by(%)		FC		RAG	
Category	**Samples**	**FC**	**RAG**	**LLM Cor.**	**NLI Ent.**	**LLM Cor.**	**NLI Ent.**
ANI	4	25.0%	**75.0%**	50.00	25.37	**100.00**	**74.77**
CNS	47	27.7%	**61.7%**	67.55	41.94	**72.87**	**53.58**
SEE	93	**45.2%**	**45.2%**	**75.54**	**61.76**	74.73	56.74
TLR	41	**53.7%**	36.6%	**63.41**	**63.80**	60.37	51.13
Total	185	42.2%	**48.1%**	70.27	**56.39**	**71.62**	55.08

Table 6: FC vs RAG across question categories, with metric breakdown, on EHRNoteQA (8K+).

5.3 Metrics and Insights

Metric Comparisons: Fig. 4 of model performance across metrics shows that instruct models lead on the single-note task (EHR-DS-QA) across semantic and NLI metrics, while larger reasoning models lead on the multi-note reasoning task (EHRNoteQA) based on LLM-as-a-judge metrics[4]. Significance testing in Table 18 confirms that the reasoning model, QwQ-32B, significantly outperforms Qwen2.5-32B on LLM Correctness for multi-note, but not on NLI or BioClinical F1, and not on any metric for the single-note. Reasoning models give factually consistent answers that go beyond the gold reference. LLM Correctness rewards this, as its rubric permits compatible additional information, whereas NLI and BioClinical F1 demand stricter reference alignment and penalize the same answers, highlighting an evaluation gap.

Metric Insights: A qualitative analysis of metric disagreements reveals three common failure modes: (1) models adding unsupported details, inflating Correctness over Faithfulness; (2) correct answers with low lexical overlap penalised by METEOR/BERTScore and (3) clinically aligned answers flagged by NLI due to negation or phrasing differences. These point to the need for more reasoning-appropriate metrics (§Appendix F.2).

6 Conclusion

We comprehensively evaluate LLM Long-Context (LC) comprehension on longitudinal, patient-centric EHRs across multiple datasets, models, and metrics under both Full-Context (FC) and RAG settings. Feeding only the most relevant content outperforms its full-notes counterpart, showing that LLMs struggle to identify the most relevant information in LC settings. While reasoning models gain on complex multi-note cases, they fail to generate factually consistent and compatible answers that go beyond the gold reference. Medically fine-tuned models do not exhibit superior performance despite their promise. Hybrid RAG with re-ranking outperforms FC, with chunk-based RAG benefiting single-note and hierarchical RAG benefiting complex multi-note reasoning and smaller models benefiting most. A category-level analysis confirms RAG's advantage in clinical nuance and absence-of-information recognition, while FC retains an advantage for longitudinal reasoning.

[4]All- and long-context metric correlations in App.F.1

Limitations

Our work focuses on MIMIC-derived datasets, which represent English-only data from a single U.S. medical center. As such, performance may not generalize to other languages, populations, and healthcare systems. Furthermore, the de-identified nature of the data limits us from examining potential cultural and other biases of LLMs over long contexts. Importantly, although our work presents an early step towards benchmarking long-context LLM performance in longitudinal patient-centric settings, our findings are intended solely for research purposes. Results reflect model performance on question answering benchmarks and should not be interpreted as guarantees of clinical safety, equitable performance, or readiness for clinical deployment. Beyond the complexities of modeling clinical notes, real-world records comprise heterogeneous data from other sources and modalities. Our work is therefore limited in its assessment of long-context by focusing only on the textual modality, something we aim to address in future work. Our analysis also highlights limitations in current evaluation metrics, specifically regarding the assessment of reasoning, the lack of transparency in LLM-as-a-judge frameworks, and the difficulty of accurately measuring faithfulness when benchmarks are supplemented with additional information. Finally, while we focus on Qwen-based models due to their competitive performance, we acknowledge that an evaluation of a wider range of LLMs could offer a more complete picture of the LLM landscape in long-form medical QA.

Ethics Statement

This work uses MIMIC-III, MIMIC-IV, and four derivative datasets (EHRNoteQA, EHR-DS-QA, RadQA, CliniQG4QA) accessed under the PhysioNet Credentialed Health Data License 1.5.0. All datasets contain de-identified patient records in accordance with HIPAA Safe Harbor standards. We exclusively used open-weight LLMs to ensure no patient data were transmitted to third-party proprietary systems.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible Ai UK (KP0016) as a Keystone project led by Maria Liakata, also supporting Talia Tseriotou and Jenny Chim.

The authors would like to thank Prof. Dr. Ferdi Sönmez for his guidance and the anonymous reviewers for their valuable feedback. Special thanks to the creators and maintainers of the MIMIC databases for their commitment to open science.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):511.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. 2025. Longhealth: A question answering benchmark with long clinical documents. *Journal of Healthcare Informatics Research*, pages 1–17.

Dina Albassam, Adam Cross, and Chengxiang Zhai. 2025. Leveraging llms for predicting unknown diagnoses from clinical notes. *arXiv preprint arXiv:2503.22092*.

Andrei Alexandru, Antonia Calvi, Henry Broomfield, Jackson Golden, Kyle Dai, Mathias Leys, Maurice Burger, Max Bartolo, Roman Engeler, Sashank Pisupati, et al. 2025. Atla selene mini: A general purpose evaluation model. *arXiv preprint arXiv:2501.17195*.

Feras AlMannaa, Talia Tseriotou, and Maria Liakata. 2026. Optimed at archehr-qa 2026: Gepa prompt optimization and multi-agent majority voting for ehr-grounded question answering. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC 2026*, pages 539–550, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

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

Anthropic. 2023. Model card and evaluations for claude models. *https://www-files.anthropic.com/production/images/Model-Card-Claude-2. pdf*.

Anna Arias-Duart, Pablo Agustin Martin-Torres, Daniel Hinjos, Pablo Bernabeu-Perez, Lucia Urcelay Ganzabal, Marta Gonzalez Mallo, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Sergio Alvarez-Napagao,

and Dario Garcia-Gasulla. 2025. Automatic evaluation of healthcare LLMs beyond question-answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 108–130, Albuquerque, New Mexico. Association for Computational Linguistics.

Manoj Ghuhan Arivazhagan, Lan Liu, Peng Qi, Xinchi Chen, William Yang Wang, and Zhiheng Huang. 2023. Hybrid hierarchical retrieval for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10680–10689, Toronto, Canada. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664, Vienna, Austria. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Charlotte M. H. H. T. Bootsma-Robroeks, Jessica D. Workum, Stephanie C. E. Schuit, Anne Hoekman, Tarannom Mehri, Job N. Doornberg, Tom P. van der Laan, and Rosanne C. Schoonbeek. 2025. Ai-generated draft replies to patient messages: exploring effects of implementation. *Frontiers in Digital Health*, Volume 7 - 2025.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025a. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025b. Towards medical complex reasoning with LLMs through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria. Association for Computational Linguistics.

Pei Chen, Hongye Jin, Cheng-Che Lee, Rulin Shao, Jingfeng Yang, Mingyu Zhao, Zhaoyu Zhang, Qin Lu, Kaiwen Men, Ning Xie, Huasheng Li, Bing Yin, Han Li, and Lingyun Wang. 2025c. LongLeader: A comprehensive leaderboard for large language models in long-context scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8734–8750, Albuquerque, New Mexico. Association for Computational Linguistics.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023a. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023b. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Amin Dada, Osman Alperen Koraş, Marie Bauer, Jean-Philippe Corbeil, Amanda Butler Contreras, Constantin Marc Seibold, Kaleb E Smith, Julian

Friedrich, and Jens Kleesiek. 2025. Does biomedical training lead to better medical performance? In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 46–59, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Pritam Deka, Anna Jurek-Loughrey, and Deepak P. 2023. Multiple evidence combination for fact-checking of health-related information. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 237–247, Toronto, Canada. Association for Computational Linguistics.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. BAMBOO: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2086–2099, Torino, Italia. ELRA and ICCL.

Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C Adams, et al. 2024. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *arXiv preprint arXiv:2408.13833*.

Yongqi Fan, Hongli Sun, Kui Xue, Xiaofan Zhang, Shaoting Zhang, and Tong Ruan. 2025a. MedOdyssey: A medical domain benchmark for long context evaluation up to 200K tokens. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 32–56, Albuquerque, New Mexico. Association for Computational Linguistics.

Yongqi Fan, Nan Wang, Kui Xue, Jingping Liu, and Tong Ruan. 2025b. MedEureka: A medical domain benchmark for multi-granularity and multi-data-type embedding-based retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2825–2851, Albuquerque, New Mexico. Association for Computational Linguistics.

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 Gatidis, Scott Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. 2024. Medalign: a clinician-generated dataset for instruction following with electronic medical records. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2288–2292, New York, NY, USA. Association for Computing Machinery.

Patricia Garcia, Stephen P. Ma, Shreya Shah, Margaret Smith, Yejin Jeong, Anna Devon-Sand, Ming Tai-Seale, Kevin Takazawa, Danyelle Clutter, Kyle Vogt, Carlene Lugtu, Matthew Rojo, Steven Lin, Tait Shanafelt, Michael A. Pfeffer, and Christopher Sharp. 2024. Artificial intelligence–generated draft replies to patient inbox messages. *JAMA Network Open*, 7(3):e243201–e243201.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William 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.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

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

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Praveen K Kanithi, Clément Christophe, Marco AF Pimentel, Tathagata Raha, Nada Saadi, Hamza Javed, Svetlana Maslenkova, Nasir Hayat, Ronnie Rajan, and Shadab Khan. 2024. Medic: Towards a comprehensive framework for evaluating llms in clinical applications. *arXiv preprint arXiv:2409.07314*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. pages 4334–4353.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Konstantin Kotschenreuther. 2024. Ehr-ds-qa: A synthetic qa dataset derived from medical discharge summaries for enhanced medical information retrieval systems. *PhysioNet*.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwang Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. *Advances in Neural Information Processing Systems*, 37:124575–124611.

Dohyeon Lee, Seung-won Hwang, Kyungjae Lee, Seungtaek Choi, and Sunghyun Park. 2023. On complementarity objectives for hybrid retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13357–13368, Toronto, Canada. Association for Computational Linguistics.

Eric Lehman, Vladislav Lialin, Katelyn Edelwina Legaspi, Anne Janelle Sy, Patricia Therese Pile, Nicole Rose Alberto, Richard Raymund Ragasa, Corinna Victoria Puyat, Marianne Katharina Taliño, Isabelle Rose Alberto, Pia Gabrielle Alfonso, Dana Moukheiber, Byron Wallace, Anna Rumshisky, Jennifer Liang, Preethi Raghavan, Leo Anthony Celi, and Peter Szolovits. 2022. Learning to ask like a physician. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 74–86, Seattle, WA. Association for Computational Linguistics.

Quinn Leng, Jacob Portes, Sam Havens, Matei Zaharia, and Michael Carbin. 2024. Long context rag performance of large language models. *arXiv preprint arXiv:2411.03538*.

Hui Yi Leong, Yifan Gao, and Shuai Ji. 2024. A gen ai framework for medical note generation. In *2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 423–429.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024a. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333, Bangkok, Thailand. Association for Computational Linguistics.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024b. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 881–893, Miami, Florida, US. Association for Computational Linguistics.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Man Luo, Arindam Mitra, Tejas Gokhale, and Chitta Baral. 2022. Improving biomedical information retrieval with neural retrievers. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11038–11046.

Kathryn A Martinez, Rebecca Schulte, Michael B Rothberg, Maria Charmaine Tang, and Elizabeth R Pfoh. 2024. Patient portal message volume and time spent on the ehr: an observational study of primary care clinicians: Martinez et al. *Journal of General Internal Medicine*, 39(4):566–572.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. 2025. Nolima: Long-context evaluation beyond literal matching. In *Forty-second International Conference on Machine Learning*.

Skatje Myers, Timothy A Miller, Yanjun Gao, Matthew M Churpek, Anoop Mayampurath, Dmitriy Dligach, and Majid Afshar. 2025. Lessons learned on information retrieval in electronic health records: a comparison of embedding models and pooling strategies. *Journal of the American Medical Informatics Association*, 32(2):357–364.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Martin Krallinger, Miguel Rodríguez-Ortega, Eduard Rodriguez-López, Natalia Loukachevitch, Andrey Sakhovskiy, Elena Tutubalina, Dimitris Dimitriadis, Grigorios Tsoumakas, George Giannakoulas, Alexandra Bekiaridou, Athanasios Samaras, Giorgio Maria Di Nunzio, Nicola Ferro, Stefano Marchesin, Marco Martinelli, Gianmaria Silvello, and Georgios Paliouras. 2025. Overview ofnbsp;bioasq 2025: The thirteenth bioasq challenge onnbsp;large-scale biomedical semantic indexing andnbsp;question answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9–12, 2025, Proceedings*, page 173–198, Berlin, Heidelberg. Springer-Verlag.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Ankit Pal and Malaikannan Sankarasubbu. 2024. Gemini goes to Med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 21–46, Mexico City, Mexico. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. pages 2357–2368.

Jie Pan, Seungwon Lee, Cheligeer Cheligeer, Elliot A. Martin, Kiarash Riazi, Hude Quan, and Na Li. 2025. Integrating large language models with human expertise for disease detection in electronic health records. *Comput. Biol. Med.*, 191(C).

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. BioMRC: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149, Online. Association for Computational Linguistics.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Jessica Ryan, Alexander I Gumilang, Robert Wiliam, and Derwin Suhartono. 2026. Self-medrag: a self-reflective hybrid retrieval-augmented generation framework for reliable medical question answering. *arXiv preprint arXiv:2601.04531*.

Saeel Sandeep Nachane, Ojas Gramopadhye, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. 2024. Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering. *arXiv e-prints*, pages arXiv–2403.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, et al. 2025. Reasonir: Training retrievers for reasoning tasks. *arXiv preprint arXiv:2504.20595*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2025. Rationale-guided retrieval augmented generation for medical question answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12739–12753, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiayu Song, Jenny Chim, Adam Tsakalidis, Julia Ive, Dana Atzil-Slonim, and Maria Liakata. 2024. Combining hierachical VAEs with LLMs for clinically meaningful timeline summarisation in social media. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14651–14672, Bangkok, Thailand. Association for Computational Linguistics.

Sarvesh Soni, Soumya Gayen, and Dina Demner-Fushman. 2025. Overview of the ArchEHR-QA 2025 shared task on grounded question answering from electronic health records. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 396–405, Viena, Austria. Association for Computational Linguistics.

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*, pages 6250–6259, Marseille, France. European Language Resources Association.

Linus Stuhlmann, Michael Alexander Saxer, and Jonathan Fürst. 2025. Efficient and reproducible biomedical question answering using retrieval augmented generation. In *2025 IEEE Swiss Conference on Data Science (SDS)*, pages 154–157. IEEE.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Liu Haisu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik, Danqi Chen, and Tao Yu. 2025. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *The Thirteenth International Conference on Learning Representations*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

Simon Šuster and Walter Daelemans. 2018. CliCR: a dataset of clinical case reports for machine reading comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1551–1563, New Orleans, Louisiana. Association for Computational Linguistics.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Qwen Team et al. 2024b. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Wenxuan Wang, Zizhan Ma, Guo Yu, Yiu-Fai Cheung, Meidan Ding, Jie Liu, Wenting Chen, and Linlin Shen. 2026. Beyond the leaderboard: Rethinking medical benchmarks for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43078–43123, San Diego, California, United States. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2024. Augmenting black-box LLMs with medical textbooks for biomedical question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1754–1770, Miami, Florida, USA. Association for Computational Linguistics.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. 2024a. Chatqa 2: Bridging the gap to proprietary llms in long context and rag capabilities. *arXiv preprint arXiv:2407.14482*.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May Dongmei Wang, Joyce C. Ho, Chao Zhang, and Carl Yang. 2024b. BMRetriever: Tuning large language models as better biomedical text retrievers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22234–22254, Miami, Florida, USA. Association for Computational Linguistics.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin

Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Tan Yu, Anbang Xu, and Rama Akkiraju. 2024. In defense of rag in the era of long-context language models. *arXiv preprint arXiv:2409.01666*.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*.

Xiang Yue, Xinliang Frederick Zhang, Ziyu Yao, Simon Lin, and Huan Sun. 2021. Cliniqg4qa: Generating diverse questions for domain adaptation of clinical question answering. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 580–587. IEEE.

Lei Zhang, Yunshui Li, Ziqiang Liu, Jiaxi Yang, Junhao Liu, Longze Chen, Run Luo, and Min Yang. 2024a. Marathon: A race through the realm of long context with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5201–5217, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024b. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, Bangkok, Thailand. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 4442–4457, New York, NY, USA. Association for Computing Machinery.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, Online. Association for Computational Linguistics.

A Inference Setup

- **Inference Engine:** Nvidia TensorRT.
- **Quantization:** FP8.
- **Hardware:** 7B parameter models: Deployed on single Nvidia H100 GPUs with 40GB VRAM. 32B parameter models with extended context: Deployed on dual Nvidia H100 GPUs, each equipped with 80GB VRAM. A total of 260 GPU hours was spent.
- **Libraries:** Qdrant, DSPy, Litellm, pandas, Evaluate, Fastembed, Pylate

B Hyperparameters

Hyperparameter		Value
LLM	temperature	inst: 0, reas: 1
	freq_penalty	0
	pres_penalty	0
	think tokens	QwQ: $20k$ HuatuoGPT: $8k$
RAG	top_k	$2\times$ num chunks
	k_{RRF}	60 (recommended by Qdrant)
	k_1	TF saturation
	b	length norm.
	avgdl	avg. doc length

Table 7: Hyperparameter Configuration

C Datasets

C.1 Context Data Distributions

Token-length distributions per dataset are shown in Figures 5 and 6.

C.2 Context Data Preparation

We performed minimal cleaning and standardization to support retrieval and prompting, including merging notes by patient and stay, normalizing timestamps, adding note-type metadata, computing token counts for context segmentation, and filtering to human-verified subsets where applicable.

We describe the cleaning and normalization steps applied prior to indexing and prompting:

- Cleaning and de-duplication of clinical notes; removal of template boilerplate where applicable.

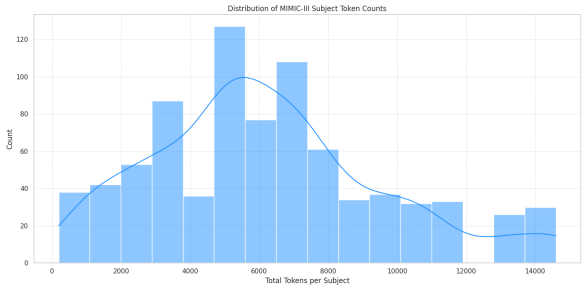


Figure 5: MIMIC-III Context Distributions

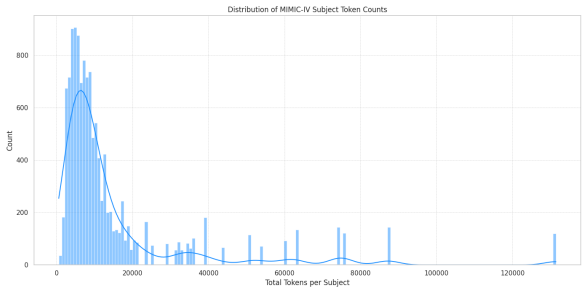


Figure 6: MIMIC-IV Context Distributions

- Merging notes by patient and hospital stay; preserving note-type and section headers for downstream retrieval.
- Datetime standardization and chronological ordering; normalization where timezones or partial timestamps occur.
- Metadata augmentation (e.g., note type, encounter identifiers) to support Include DS vs Include All settings.
- Tokenization and token-count computation at note- and patient-level for context segmentation.

These details enable reproducibility of sampling and retrieval segmentation.

	0–8K (Include All)	8–16K (Include All)	16–32K (Include All)	32–128K (Include All)
Include Related	2,247	4,767	5,149	28,131
Exclude Related	1,284	5,712	13,733	24,472
Include All	3,934	9,851	18,909	41,409

Table 8: Context Length per Formulation on the combined EHR-DS-QA and EHRNoteQA datasets.

C.3 Context Length per Formulation

We provide the Median token context length per original Full-Context (Include All) bin based on the combined EHR-DS-QA and EHRNoteQA datasets on which we evaluate the open-ended generative

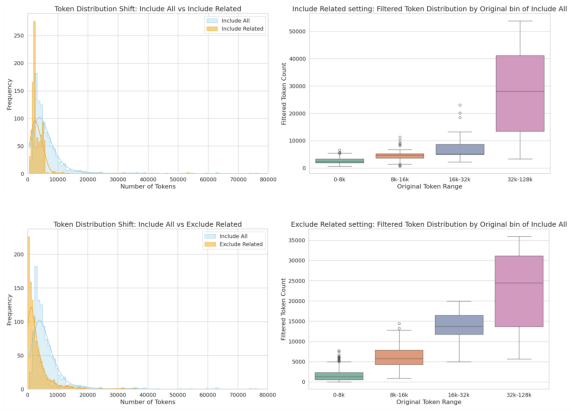


Figure 7: Distribution of tokens of Include Related versus Include All and Exclude Related versus Include All on the combined EHR-DS-QA and EHRNoteQA datasets.

Dataset	LLM-as-a-judge	QA Pairs
EHR-DS-QA	Selene-8B	113
	Prometheus-8x7B-v2.0	98
	Qwen2.5-32B-Instruct	71
EHRNoteQA	Selene-8B	180
	Prometheus-8x7B-v2.0	220
	Qwen2.5-32B-Instruct	42

Table 9: Disagreement instances where NLI Med Contradiction is high (> 0.7) and LLM Correctness is also high (> 0.7).

QA capabilities. Table 8 shows that as expected for each of the original bins the median lengths fall compared to the *Include All* setting in both the *Include Related* and *Exclude Related* settings, with the most drastic change being in the *Include Related* setting context-bin of 16-32K tokens from a median of $\sim$19K to a median of around $\sim$5K tokens. Additionally,in Figure 7 we also provide the distribution of tokens of *Include Related* versus *Include All* and *Exclude Related* versus *Include All*.

C.4 Prompt Formulations

We used a simple zero-shot prompt structure and tailored it explicitly for each task formulation:

- **Extractive**: Request the model to answer by extracting the most relevant answer from the context.
- **Multiple-choice**: Standard multiple choice prompt.
- **Open-ended**: Focusing on open ended question answering favoring short, single sentence answers.

Dataset	LLM-as-a-judge 1	LLM-as-a-judge 2	QA Pairs
EHR-DS-QA	Prometheus	Qwen2.5-32B-Instruct	520
	Prometheus	Selene-8B	363
	Qwen2.5-32B-Instruct	Selene-8B	664
EHRNoteQA	Prometheus	Qwen2.5-32B-Instruct	690
	Prometheus	Selene-8B	236
	Qwen2.5-32B-Instruct	Selene-8B	519

Table 10: Judge-pair disagreements where the absolute difference in LLM Correctness ≥ 0.5.

Metric	Splade	BM25
LLM Correctness	67.30	**69.34**
LLM Completeness	65.41	**67.14**
LLM Faithfulness	**47.64**	46.23
NLI Med Entailment	52.68	**54.35**
NLI Med Contradiction	16.98	**16.94**
Bio BERTScore F1	**79.75**	79.67
METEOR	**43.93**	43.63

Table 11: Comparison of Different Sparse Retrievers

Below is the sample prompt for extractive tasks:

```
System message:

Your input fields are:
1. `medical_record` (list[str]): List of patient notes (chronological order).
2. `question` (str): A question about the patient's record.
Your output fields are:
1. `answer` (str): Short single-sentence answer to the question.

All interactions will be structured in the following way, with the appropriate values filled in.

[[ ## medical_record ## ]]
{medical_record}

[[ ## question ## ]]
{question}

[[ ## answer ## ]]
{answer}

[[ ## completed ## ]]

In adhering to this structure, your objective is:
Given a patient's medical record and a question, answer the question correctly and shortly.
```

C.5 Dataset Comparisons

Aiming to stress-test LLM capabilities under real patient-centric scenarios, our dataset selection process was based on a extensive grid of datasets. Table 12 summarizes the QA datasets we analyzed in our selection process, including the EHR candidates. The chosen EHR-based, human-verified datasets provide diverse but comparable settings across generative, extractive, and MC formulations. While some exceed 8K tokens, supporting our long-context evaluation, they all provide strong augmentation potential towards a multi-note longer evaluation setting that allows content extension up to 128K for each patient (§3.1), enabling a realistic comparison of Full-context and RAG pipelines. We provide citations of each considered dataset from Table 12 in Table 13.

D Model Comparisons

D.1 Sparse Retrievers

On Table 11 we provide a sparse retriever comparison on the EHR-DS-QA dataset using the Qwen2.5-7B-Instruct-1M LLM and Qwen3-Embedding-8B as the dense retriever. BM25 has better performance across LLM Correctness, LLM Completeness and NLI-based metrics while showcasing slightly inferior performance on BioClinical BERTScore F1, METEOR and LLM Faithfulness.

D.2 Chunking Strategy

During the development phase of the RAG approach we experiment with different chunking strategies for long-context of 8K+ tokens, comparing the retrieval of 512-token chunks, 1K-token chunks with a 12-token overlap, and full notes, all indexed and retrieved with the same dense retriever. We present the results in Table 14 and note that across context size bins of 8–16K, 16–32K and 32–128K, the 512 chunking setting has the best performance across metrics with the exception of NLI metrics in the 16–32K bin.

D.3 LLM-as-a-judge

In Table 9 we examine the number of QA cases for each dataset that the different Judges cause disagreement with the NLI Med Contradiction metric. We observe that Qwen2.5-32B-Instruct [5] has far less such disagreement compared to more specialized judges. Table 10 also provides direct comparisons of LLM-as-a-judge models, demonstrating cases where LLM Correctness between such pairs is more than 0.5 and therefore high. Selene-8B [6] and Prometheus-8x7B-v2.0 [7] comparisons show that these two models are consistently more in agreement between them while Qwen2.5-32B-Instruct shows more independent judging abilities.

D.4 Context Size Performance

Figure 8 presents the performance of each model on open-ended generative QA across three different metrics.

E Statistical Significance Tests

We performed Philipp Koehn's Koehn (2004) paired bootstrap resampling significance tests

[5] https://huggingface.co/Qwen/Qwen2.5-32B-Instruct

[6] https://huggingface.co/AtlaAI/Selene-1-Mini-Llama-3.1-8B

[7] https://huggingface.co/prometheus-eval/prometheus-8x7b-v2.0

Dataset	# QA Pairs	Token Len. (max)	Task Types	Synthetic	Annotation	Reasoning
Medical Textbooks, Websites, Knowledge Bases						
MedQuAD	47,457	< 2K	extractive		automatic	✗
MashQA	34,808	< 2K	extractive (multi-span)		experts	✓
MedOdyssey (En.KG)	100	< 128K	extrative (graph)		model&human	✓
MedQA (USMLE)	12,723	> 1M	MC		experts	✓
Biomedical Literature						
BioMRC (LARGE)	812,707	< 2K	MC, extractive		automatic	✗
BioASQ (b)	5,729	< 8K	extractive, generative		experts	✓
PubMedQA	1,000	< 2K	MC, generative		experts	✓
Clinical Case Reports						
CliCR	104,919	< 8K	extractive (cloze)		automatic	✓
Patient Notes						
LongHealth	400	< 8K	MC	✓	experts	✓
DiSCQ	2,029	< 2K	extractive, generative		experts	✓
RadQA	6,148	< 16K	extractive		experts	✓
CliniQG4QA	8,824/1,287	< 8K	extractive		model/experts	✗
EHR-DS-QA	156,599/478	< 8K	generative	✓	model/experts	✗
EHRNoteQA	962	< 8K	MC, generative		experts	✓

Table 12: Biomedical Datasets Comparison. References for each dataset are in Table 13.

Dataset	Source
MedQA	Jin et al. (2021)
MedQuAD	Abacha and Demner-Fushman (2019)
MashQA	Zhu et al. (2020)
MedOdyssey	Fan et al. (2025a)
BioMRC	Pappas et al. (2020)
BioASQ	Nentidis et al. (2025)
PubMedQA	Jin et al. (2019)
CliCR	Šuster and Daelemans (2018)
LongHealth	Adams et al. (2025)
MediNote	Leong et al. (2024)
DiSCQ	Lehman et al. (2022)
RadQA	Soni et al. (2022)
CliniQG4QA	Yue et al. (2021)
EHR-DS-QA	Kotschenreuther (2024)
EHRNoteQA	Kweon et al. (2024)

Table 13: Corresponding Citations for Datasets considered in Dataset Selection of Table 12

across experiments to examine the significance of our findings, based on 5,000 repetitions. For each comparison we bootstrap-resample questions with replacement, keeping system scores paired within each resample, and report the fraction of resamples in which the first system's mean does not exceed the second's as a one-sided p-value ($p < 0.05$ as significant). Being paired and distribution-free, the test directly addresses whether one system is significantly better on the same questions, without normality assumptions.

In this section we present our results with a corresponding table and discussion per Claim. Claims 1–4 use the full evaluation set over all four context bins, in the *Include All* setting (Claims 1 and 4), the *Exclude All* setting (Claim 2) and *Exclude Related* versus *Exclude All* (Claim 3). Claim 5 uses the 8K+ bins of the open-ended datasets; each question is scored by the mean of LLM Correctness, NLI Entailment and BioClinical F1, and the RAG setting per model is selected based on the best model-metric average, compared against the *Include All setting*.

Claim 1: Medically fine-tuning does not improve over the base model. Strongly supported no statistically significant differences between the medically fine-tuned HuatuoGPT-o1-7B and its instruction-tuned base Qwen2.5-7B across metrics on either of the open-ended datasets: EHRNoteQA and EHR-DS-QA (Table 15). We therefore describe this as an absence of measurable benefit rather than as a degradation.

Claim 2: Smaller models exhibit stronger memorization than larger ones. Strongly supported that Qwen2.5 7B significantly outperforms Qwen2.5 32B in the *Exclude All* setting on five of the six dataset-metric pairs, the exception being BioClinical F1 on EHR-DS-QA (Table 16).

Claim 3: Supplementary material is informa-

Setting	Token Range	LLM Corr.	LLM Comp.	LLM Faith.	NLI Entail.	NLI Contra.	Bio Clin. F1	METEOR
RAG 512		**41.52**	**35.27**	**40.18**	**50.50**	**18.55**	**77.69**	**35.01**
RAG 1K	8-16K	38.84	33.04	35.71	36.74	19.72	76.90	32.65
RAG Full Note		39.73	28.57	36.61	43.71	19.84	75.49	33.31
RAG 512		**60.85**	**54.72**	**51.42**	55.75	17.37	**80.49**	**45.11**
RAG 1K	16-32K	53.77	47.17	50.94	**61.68**	**13.65**	79.44	41.87
RAG Full Note		55.19	47.64	45.28	60.17	13.66	78.60	41.80
RAG 512		**52.62**	**41.28**	**43.31**	**55.90**	**18.45**	**79.80**	**42.06**
RAG 1K	32-128K	46.51	38.37	40.99	49.24	21.02	79.14	38.98
RAG Full Note		50.87	37.79	41.57	55.29	22.77	79.10	41.58

Table 14: RAG chunk size ablation across token ranges and metrics for the EHR-DS-QA dataset using `Qwen2.5-7B-Instruct-1M`.

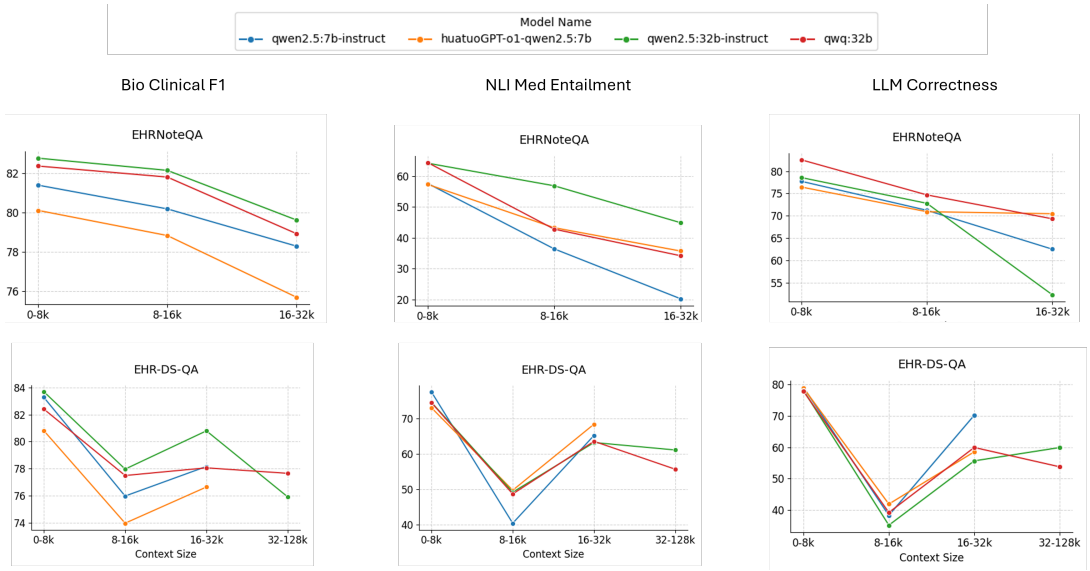


Figure 8: Model performance over context size across datasets and metrics.

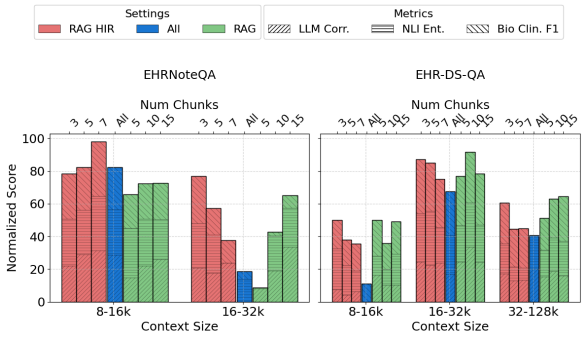


Figure 9: RAG performance across settings with Qwen2.5:32B-Instruct, min-max normalized across LLM Correctness, NLI Entailment and Bio F1. The underlying numbers are the Qwen2.5-32B rows of Table 3.

Dataset	Metric	HuatuoGPT mean	Qwen2.5 7B mean	p-value	$p < 0.05$
EHRNoteQA	LLM Correctness	0.7539	0.7630	0.7936	✗
	NLI Entailment	0.5414	0.5294	0.2184	✗
	BioClinical F1	0.7979	0.8111	1.0000	✗
EHR-DS-QA	LLM Correctness	0.6993	0.7021	0.5690	✗
	NLI Entailment	0.6558	0.6859	0.8980	✗
	BioClinical F1	0.7892	0.8075	0.9996	✗

Table 15: Paired bootstrap tests comparing the medically fine-tuned HuatuoGPT-o1-7B with the instruction-tuned base Qwen2.5-7B in the *Include All* setting, across the open-ended generative datasets and metrics.

tive for generative tasks. Strongly supported that *Exclude Related* significantly outperforms *Exclude All* on five of the six dataset-metric pairs of the open-ended generative benchmarks, the exception being NLI Entailment on EHR-DS-QA (Table 17).

Claim 4: Reasoning models offer an advantage in complex multi- versus single- note open-ended QA on LLM Correctness but not on NLI and BioClinical metrics. Supported with

Dataset	Metric	Qwen2.5 7B mean	Qwen2.5 32B mean	p-value	$p < 0.05$
EHRNoteQA	LLM Correctness	0.0699	0.0133	0.0000	✓
	NLI Entailment	0.2242	0.1090	0.0000	✓
	BioClinical F1	0.6784	0.6658	0.0000	✓
EHR-DS-QA	LLM Correctness	0.2284	0.1811	0.0002	✓
	NLI Entailment	0.3264	0.1910	0.0000	✓
	BioClinical F1	0.7222	0.7225	0.5440	×

Table 16: Paired bootstrap tests on memorization, comparing Qwen2.5 7B with Qwen2.5 32B in the *Exclude All* setting across open-ended generative datasets and metrics.

Dataset	Metric	Excl. Related mean	Excl. All mean	p-value	$p < 0.05$
EHRNoteQA	LLM Correctness	0.1968	0.0133	0.0000	✓
	NLI Entailment	0.1685	0.1090	0.0000	✓
	BioClinical F1	0.7212	0.6658	0.0000	✓
EHR-DS-QA	LLM Correctness	0.2187	0.1811	0.0148	✓
	NLI Entailment	0.2072	0.1910	0.1784	×
	BioClinical F1	0.7272	0.7225	0.0300	✓

Table 17: Paired bootstrap tests comparing *Exclude Related* with *Exclude All* for Qwen2.5 32B, assessing whether supplementary patient material is informative, across open-ended generative datasets and metrics.

QwQ 32B significantly outperforming Qwen2.5 32B on LLM Correctness only for EHRNoteQA (multi-note), but not on any metric on EHR-DS-QA (single-note), the single-note dataset (Table 18). This suggests that reasoning models generate factually consistent and compatible answers that go beyond the gold reference, which LLM Correctness rewards as it permits compatible additional information, whereas NLI and BioClinical F1 penalize the same answers because of their stricter reference alignment requirements. This further supports our argument that more reasoning-appropriate metrics are needed (§5.3).

Dataset	Metric	QwQ 32B mean	Qwen2.5 32B mean	p-value	$p < 0.05$
EHRNoteQA	LLM Correctness	0.8095	0.7701	0.0000	✓
	NLI Entailment	0.5972	0.6269	0.9574	×
	BioClinical F1	0.8218	0.8257	0.9980	×
EHR-DS-QA	LLM Correctness	0.6661	0.6678	0.5372	×
	NLI Entailment	0.6647	0.6752	0.6888	×
	BioClinical F1	0.8038	0.8114	0.9706	×

Table 18: Paired bootstrap tests comparing the QwQ 32B reasoning and Qwen2.5 32B models under the *Include All* setting, across open-ended generative datasets and metrics.

Claim 5: RAG outperforms FC on single-note QA, with hierarchical RAG superior for multi-note reasoning in smaller models. Supported through testing on 8K+ context and selecting the best RAG configuration per dataset and model by metric average against *Include All* (Table 19). RAG significantly outperforms *Include All* on the single-note EHR-DS-QA across all three models. For the more complex multi-note EHRNoteQA, significant gains are observed for the smaller 7B models, while Qwen2.5 32B shows no significant improvement under either open-ended or multiple-choice formulations, suggesting that larger models benefit less from retrieval in complex multi-note reasoning settings.

Dataset	Model/Setting	RAG avg	FC avg	p-value	p<0.05
Open-ended					
EHRNoteQA	HuatuoGPT-o1 7B (RAG HIR 5)	0.6863	0.6252	0.0016	✓
	Qwen2.5 7B (RAG HIR 5)	0.6882	0.6023	0.0000	✓
	Qwen2.5 32B (RAG HIR 3)	0.6897	0.6943	0.5998	×
EHR-DS-QA	HuatuoGPT-o1 7B (RAG 15)	0.7024	0.6193	0.0040	✓
	Qwen2.5 7B (RAG 10)	0.6748	0.6164	0.0110	✓
	Qwen2.5 32B (RAG 10)	0.6703	0.6275	0.0132	✓
Multiple-Choice					
EHRNoteQA	HuatuoGPT-o1 7B (RAG HIR 7)	0.6718	0.6252	0.0056	✓
	Qwen2.5 7B (RAG HIR 7)	0.6813	0.6023	0.0000	✓
	Qwen2.5 32B (RAG HIR 7)	0.7132	0.6943	0.1306	×

Table 19: Paired bootstrap tests for the best on average RAG strategy per dataset and model (using metric average) versus Full-Context (FC) in the *Include All* setting for 8K+ context tokens.

F Metrics

F.1 Correlations

We examine correlations between evaluation metrics to identify metric independence. On EHRNoteQA and EHR-DS-QA across all non-exclude settings (Figure 10) as well as across long-context high LLM Correctness non-exclude settings (Figure 11), LLM-as-a-judge evaluation metrics exhibit strong inter-correlations while METEOR and BERTScore are highly correlated. Although the above correlations remain high, they are less pronounced for the long-context high LLM Correctness setting of Figure 11. In general all metrics are less correlated for the long-context high LLM Correctness setting.

F.2 Blind Spots

Table 20 showcases a qualitative view of what our metrics are capturing and missing, based on an analysis of disagreement on the open-ended generative QA benchmarks.

G Case Studies

Here we show some sample studies of our error analysis.

EHR-DS-QA: Case 10083814 (final diagnoses). Three discharge summaries exist across distinct admissions; each lists diagnoses. The correct target

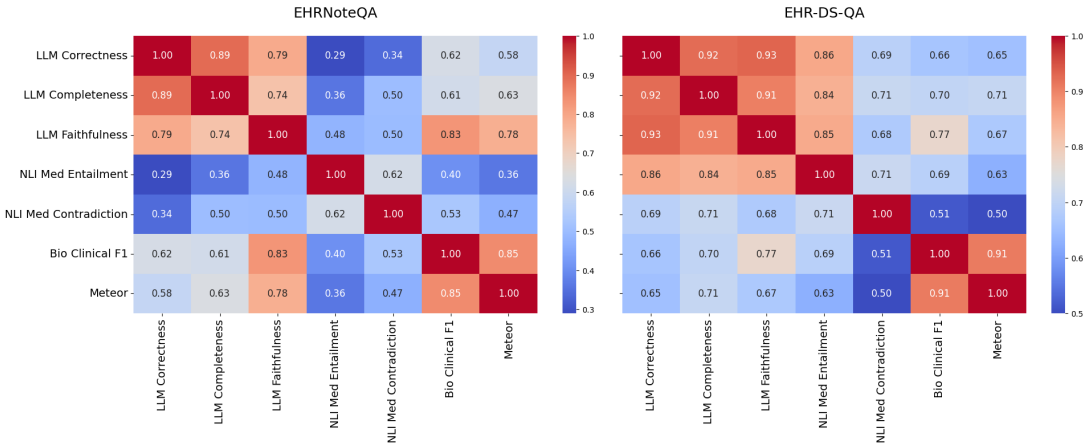


Figure 10: Metric correlations across all non-exclude settings.

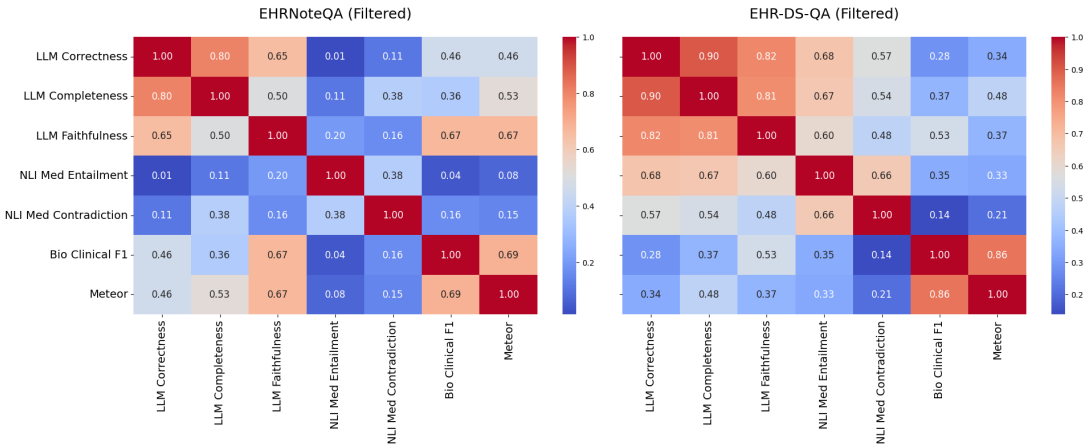


Figure 11: Metric correlations across all non-exclude settings filtered by LLM Correctness > 50 and Context $>= 16$K.

What they capture well	What they miss	Typical pattern	Examples (dataset #subject)
LLM Correctness			
Whether the main claim is right.	Unsupported add-ons that don't change the core fact.	Correctness high; Faithfulness low (adds extra).	EHRNoteQA #10043423 (core claim right, adds recommendation); EHRDSQA #10094318 (chief complaint right, extra symptom).
LLM Completeness			
Coverage of the requested pieces/elements.	Penalizes omissions but not ungrounded embellishments.	Completeness high; Faithfulness low (fully covers, then embellishes).	EHRNoteQA #10043423 (covers requested items + extras); EHRDSQA #19796003 (lists largely complete; minor extras elsewhere reduce faithfulness/F1).
LLM Faithfulness			
Grounding to evidence/gold; flags hallucinated or embellished content.	Can be over-strict on benign context or plausible but unsupported expansions.	Correct/Complete high; Faithfulness low (ungrounded detail).	EHRNoteQA #10043423 (extra recommendation not supported); EHRDSQA #10131388 (adds follow-up/psychiatric context not explicit).
NLI Med Entailment / Contradiction			
Sentence-level logical relation to gold (entailed vs. contradicted).	Sensitive to negation, hedging (*"low risk"* vs *"no risk"*), and phrasing templates.	Correctness high; Entailment low or Contradiction high (wording/negation mismatch).	EHRNoteQA #10494486 (mechanism phrasing hurts entailment); EHRDSQA #10010655 (*"no risk"* vs *"low acute risk"* triggers contradiction).
Meteor			
Surface-form overlap, good for close paraphrases.	Under-rewards long phrases, synonyms, and re-phrasings with low lexical overlap. Incorrectly rewards inaccuracies with lexical overlap.	Correctness high; Meteor low (semantic match, low surface match).	EHRDSQA #10076958 (concise "diffuse ischemic bowel" vs long gold narrative); #19796003 correct single word answer but low score; EHRNoteQA #10404814 (correct causal link, different wording/structure).
BioClinical F1			
Token/span overlap for clinical entities; good checklist signal for missing/extra items.	Under-rewards clinically acceptable reformulations (class vs. specific drug),. Incorrectly rewards inaccuracies with lexical overlap.	Correctness high; F1 moderate/low (one entity missing or formatted differently).	EHRNoteQA #10043423 (right ideas, entity list/format differences lower F1); EHRDSQA #19796003 (near-exact meds but lower F1 for small misses).

Table 20: Metric blind spots: what each metric captures and misses, based on manual analysis of metric disagreement cases, with typical patterns and representative examples from EHRNoteQA and EHR-DS-QA.

for "final diagnoses" is the most recent summary. Earlier summaries contain different diagnoses that are no longer *final* at discharge, explaining prediction–gold divergence without model hallucination.

EHRNoteQA: Case 15877599 (cause of AKI). The note supports a causal chain: gastroenteritis $\rightarrow$ increased ostomy output $\rightarrow$ severe dehydration (prerenal) $\rightarrow$ acute kidney injury. The gold captures the underlying illness; a model response may describe the immediate mechanism. Both are clinically coherent parts of the same sequence.

EHR-DS-QA: Case 10023117 (blood pressure at discharge). Predictions expressed as exact values versus ranges produce different behaviors across metrics. Ranges can be faithful to notes yet fail entailment or completeness thresholds defined against a single gold value.

H Qualitative Analysis: RAG vs FC

In Table 21 we summarize our insights about case categories where either RAG or FC tend to be more beneficial, providing specific use cases by open-ended generative dataset. We also provide examples along each of our main findings below:

- *FC performs better for questions requiring comprehensive understanding* of the entire document or when answers are located in structured sections (i.e. lists) or temporal sequences.

  Examples include:

  - Questions like "What is the patient's discharge condition?" or "What were the patient's discharge diagnoses?".
  - Temporal sequence questions, such as "What surgeries has the patient undergone and in what order?".

- *RAG outperforms FC for questions that require retrieving specific details or synthesizing scattered information* across multiple parts of the document.

  Examples include:

  - Questions like "What family history does the patient have?" or "How many tablets of dilaudid did the patient receive?".
  - Synthesis tasks, such as "What were the patient's postoperative course details?".
  - Asking about specific dates like "What was the outcome of the patient's colonoscopy as described in the discharge summary from the stay starting on 2113-09-30?" or "What was the patient's diagnosis for the hospital admission on 2154-01-28…" and other examples.

- *RAG tends to perform better overall in*

datasets where specific information is required from complex or lengthy documents.

For example:

  - In the EHRNoteQA dataset, RAG consistently outperformed FC for questions needing specific details from notes or summaries, such as "What was the outcome of the patient's colonoscopy?"

- *There is mixed evidence suggesting FC might perform better for tasks involving inferential reasoning* or identifying the absence of information.

  For example:

  - Questions like "Were there any complications during the procedure?" where RAG retrieves statements like "No complications," potentially diminishing FC's advantage.
  - Subtle inference tasks, such as "Does the patient have any psychological issues?" where FC occasionally performs better, though inconsistently.

I LLM-as-a-Judge Prompts

Below we present the three rubric prompts used for the LLM-as-a-judge evaluation (§3.4). Each call receives: i) `question`, ii) `generated answer` and iii) `reference answer`, and returns a free-text `reasoning` field followed by an integer `score` between 1–5. Each score is then mapped to $[0, 1]$ through $(s - 1)/4$. We use QWEN-2.5-32B-INSTRUCT with temperature $= 0$ (see §D.3 for model selection).

Correctness Prompt

Correctness

You are an expert medical QA evaluator. Your task is to evaluate the factual accuracy of a prediction by comparing it strictly against the ground truth reference answer.

Evaluation Process
Step 1: Identify all factual claims in the prediction
Step 2: For each claim, check if it contradicts the ground truth
Step 3: Only penalize for contradictions or factual errors, NOT for extra details

Scoring Rules
- Score 5 if ALL factual claims are accurate or neutral (no contradictions with ground truth)
- Score 4 if most claims are correct with few minor factual inaccuracies that don't change core meaning
- Score 3 if roughly half the factual claims are correct
- Score 2 for at least a single correct factual claim
- Score 1 for no correct claims at all or complete failure to answer

CRITICAL: Extra information that does not contradict ground truth = Score 5
CRITICAL: Missing factual information from ground truth = reduce score
CRITICAL: Contradicting ground truth = reduce score
CRITICAL: Adding neutral/compatible information = do NOT reduce score
CRITICAL: No correct claims at all = Score 1

Example: If the ground truth states "patient went to hospital" and the prediction states "patient went to hospital on Tuesday", this is Score 5 because adding "on Tuesday" does not contradict the ground truth.

Completeness Prompt

Completeness

You are an expert medical QA evaluator. Your task is to evaluate how thoroughly the prediction covers the clinically relevant elements present in the ground truth answer.

Evaluation Process
Step 1: List all elements/information present in ground truth
Step 2: Check if each ground truth element is covered in prediction (rewording allowed)
Step 3: Score based ONLY on coverage of ground truth elements

Scoring Rules
- Score 5 if ALL elements from ground truth are covered in prediction
- Score 4 if most elements covered with minor omissions
- Score 3 if roughly half the elements are covered
- Score 2 if few elements are covered
- Score 1 if ALL elements from ground truth are missing

CRITICAL: Extra information in prediction beyond ground truth = IGNORE completely
CRITICAL: Only count what's missing from ground truth, never penalize for extras
CRITICAL: Perfect rewording of ground truth content = Score 5
CRITICAL: All ground truth elements are missing = Score 1

Example: If ground truth says "patient has fever" and prediction says "patient has fever and headache", this is Score 5 because all ground truth content (fever) is covered. The extra "headache" is ignored for completeness scoring.

Faithfulness Prompt

Faithfulness

You are an expert medical QA evaluator. Your task is to evaluate whether every statement in the prediction is supported by established medical knowledge or clearly inferable from the ground truth.

EVALUATION FOCUS:
- Identify any unsupported claims, hallucinations, or fabricated information
- Evaluate support for each factual statement in prediction
- Ground truth serves as primary support reference - Consider established medical knowledge as secondary support

SCORING CRITERIA (1-5):
Score 5: All statements in prediction are well-supported by ground truth or established medical knowledge.
Score 4: Nearly all statements supported, with minimal harmless unsupported details.
Score 3: Roughly half supported and unsupported claims present.
Score 2: Few supported and many unsupported claims or moderate hallucinations present.
Score 1: No supported claims at all, significant hallucinations, or fabrications.

CRITICAL: Examine each factual claim in prediction for support/evidence
CRITICAL: Claims supported by ground truth or established medical facts score well
CRITICAL: Unsupported, fabricated, or speculative claims reduce faithfulness
CRITICAL: If prediction says "I don't know" when ground truth is "Cannot be answered", this is faithful (score 5)
CRITICAL: ALL claims are supported = Score 5
CRITICAL: No claim is supported at all = Score 1

Remember: Focus on identifying unsupported information, not completeness or style.

Insight	Favored	Explanation	Supporting Examples
Specific Fact Retrieval	RAG	RAG excels at extracting precise, well-defined medical facts (dates, medications, lab values, procedures) that are typically documented in structured sections of medical records.	**EHRNoteQA:** *Subject 15036658:* Colonoscopy outcome from specific date (RAG: 0.632 vs. FC: 0.399) *Subject 11049732:* Medication changes (RAG: 0.829 vs. FC: 0.352) *Subject 17818938:* Surgical procedure for erectile dysfunction (RAG: 0.918 vs. FC: 0.479) **EHR-DS-QA:** *Subject 10131388:* Dilaudid tablet count (RAG: 0.982 vs. FC: 0.778) *Subject 19926045:* DVT medication (RAG: 0.901 vs. FC: 0.564) *Subject 10090787:* Discharge medications (RAG: 0.425 vs. FC: 0.190)
Temporal Information Processing	Mixed	RAG excels at explicit temporal facts (specific dates, temporal relationships) while FC is better at temporal reasoning (sequencing, duration calculation, recognizing absence of temporal information).	**RAG Advantage:** *EHRNoteQA 15036658:* Specific date anchoring *EHRNoteQA 18467824:* Temporal relationship between admissions *EHR-DS-QA 10264949:* Nausea/vomiting timing **FC Advantage:** *EHRNoteQA 11552479:* Temporal sequencing (FC: 0.696 vs. RAG: 0.236) *EHR-DS-QA 10751849:* Duration calculation (FC: 0.541 vs. RAG: 0.320) *EHR-DS-QA 19397212:* Absence of temporal information (FC: 0.426 vs. RAG: 0.180)
Medical Terminology and Technical Content	RAG	RAG performs better with specialized medical terminology, complex procedures, and technical test results due to its ability to locate and interpret specific sections containing this information.	**EHRNoteQA:** *Subject 17445067:* Diagnosis and surgical procedure details (RAG: 0.708 vs. FC: 0.328) *Subject 18122852:* MRI and EMG test findings (RAG: 0.750 vs. FC: 0.454) *Subject 16313269:* Brain mass pathological diagnosis (RAG: 0.915 vs. FC: 0.372) **EHR-DS-QA:** *Subject 10044189:* Necrotic ulcer treatment (RAG: 0.762 vs. FC: 0.465) *Subject 19401508:* Treatment for hyponatremia (RAG: 0.479 vs. FC: 0.262)
Discharge Planning and Instructions	RAG	RAG performs better on Discharge information (instructions, medications, condition) that is usually well-structured in specific sections.	**EHRNoteQA:** *Subject 11690633:* Discharge condition and instructions (RAG: 0.762 vs. FC: 0.349) *Subject 11863782:* Discharge disposition and medications (RAG: 0.749 vs. FC: 0.327) **EHR-DS-QA:** *Subject 10921250:* Discharge condition (RAG: 0.856 vs. FC: 0.466) *Subject 10940920:* Discharge instructions (RAG: 0.650 vs. FC: 0.253) *Subject 10064678:* Discharge instructions (RAG: 0.352 vs. FC: 0.094)
Cause–Effect and Relationship Understanding	RAG	RAG is better at understanding relationships (symptom–procedure, medication–outcome, test result–action).	**EHRNoteQA:** *Subject 13032648:* Causes of leg pain and surgical procedure (RAG: 0.730 vs. FC: 0.293) *Subject 15748482:* Flomax usage reason and outcome (RAG: 0.819 vs. FC: 0.495) *Subject 17436366:* Blood/urine culture results and actions (RAG: 0.718 vs. FC: 0.445) **EHR-DS-QA:** *Subject 19926045:* Lovenox symptom management (RAG: 0.545 vs. FC: 0.334) *Subject 19401508:* Admission cause and treatment (RAG: 0.372 vs. FC: 0.108)
Holistic Patient Understanding	FC	FC excels when questions require synthesis of information across multiple document sections to build a complete picture of the patient's status, multiple diagnoses.	**EHRNoteQA:** *Subject 18753609:* Therapeutic interventions for leg pain (FC: 0.675 vs. RAG: 0.450) **EHR-DS-QA:** *Subject 10262565:* Discharge condition (FC: 0.882 vs. RAG: 0.681) *Subject 10049941:* Discharge diagnoses (FC: 0.579 vs. RAG: 0.188) *Subject 10978236:* Discharge condition – unusual circumstance (FC: 0.321 vs. RAG: 0.093)
Absence or Negative Information Recognition	FC	While findings remain inconsistent with both RAG and FC demonstrating mostly low scores, FC seems more promising at recognizing when information is absent or when negative findings are documented as it can assess the entire document context.	**EHR-DS-QA:** *Subject 19397212:* Absence of symptom presentation timing (FC: 0.426 vs. RAG: 0.180) *Subject 10751849:* Absence of major procedures (FC: 0.243 vs. RAG: 0.037) *Subject 10264949:* Absence of social factors (FC: 0.497 vs. RAG: 0.149) *Subject 19397212:* Absence of age information (FC: 0.725 vs. RAG: 0.452)

Table 21: RAG vs FC Qualitative Analysis.