Investigating LLM Capabilities on Long Context Comprehension for Medical Question Answering

Anonymous ACL submission

Abstract

This study is the first to investigate LLM comprehension capabilities over long-context (LC), clinically relevant medical Question Answering (QA) beyond MCQA. Our comprehensive approach considers a range of settings based on content inclusion of varying size and relevance, LLM models of different capabilities and a variety of datasets across task formulations. We reveal insights on model size effects and their limitations, underlying memorization issues and the benefits of reasoning models, while demonstrating the value and challenges of leveraging the full long patient's context. Importantly, we examine the effect of Retrieval Augmented Generation (RAG) on medical LC comprehension, uncovering best settings in single versus multi-document QA datasets. We shed light into some of the evaluation aspects using a multi-faceted approach uncovering common metric challenges. Our quantitative analysis reveals challenging cases where RAG excels while still showing limitations in cases requiring temporal reasoning.

1 Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Dubey et al., 2024; Yang et al., 2024) perform impressively on medical tasks (Borgeaud et al., 2021; 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., 2023; 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., 2024b; 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). Assessing LLM capabilities on context from expert-curated electronic health records (EHRs) could provide crucial signals on how well models address complexities in real data, including domain-specific vocabulary, heterogeneous document types, multi-document reasoning, data noise, linguistic diversity, long contexts, and long-range dependencies (Wornow et al., 2023). However, research on open-form QA pertaining to EHRs focuses on creating datasets (Yang et al., 2022) or EHR-specialized models (Fleming et al., 2023), rather than investigating LLM performance in real-world QA.

Moreover, most EHR-based benchmarks consider single-document contexts (Pampari et al., 2018; Yue et al., 2020), while in practice there is no guarantee that the necessary information will be included within a specific patient document. Reasoning over multiple long, temporally-dependent documents further poses questions around long-context (LC) limitations and evidence positioning sensitivity (Liu et al., 2023; Xiao et al., 2023), also observed in medically focused studies (Adams et al., 2024). Work on the effect of long-range dependencies within medical machine reading comprehension (MRC) is limited (Vatsal and Singh, 2024), focusing on shorter texts (∼1.5-4K tokens) and span-based QA rather than complex EHRs. Work on LC medical QA involves either synthetic data, is MCQA-based and still relatively short (up to 6k) (Adams et al., 2024)), or small in sample size and not human-validated (Fan et al., 2024). We address this gap by investigating LLM capabilities to answer questions given longitudinal patient-centric EHRs, where QA pairs are human-validated.

While recent LLM releases push the frontiers of context size, performance does not scale with increased context (Levy et al., 2024; Modarressi et al., 2025). Retrieval Augmented Generation (RAG) (Lewis et al., 2021) often forms a cost efficient alternative by selecting relevant context

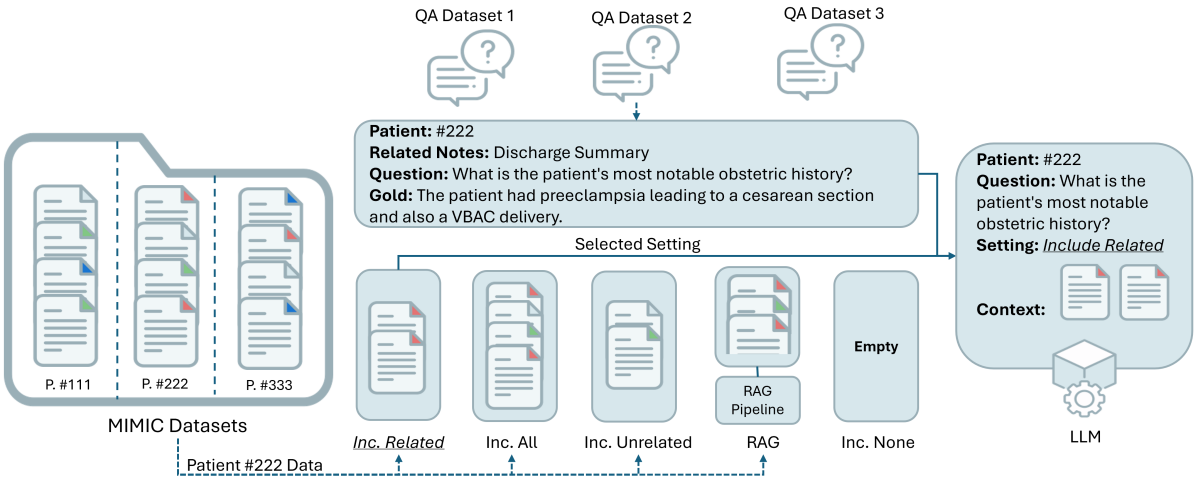

Figure 1: Overview of our Approach

on demand. What is preferable is subject to debate, with each showing benefits for different tasks and settings (Li et al., 2024b): RAG has been reported to enhance LLM performance compared to LC prompt compression for QA (Zhang et al., 2024a), outperforming SOTA LLMs on both 32K and 128K benchmarks (Xu et al., 2024a; Bai et al., 2024). By contrast, others showed that RAG performance peaks at 32K context length for large LLMs (Leng et al., 2024). While work on medical QA reports promising gains when retrieval is carefully designed (Xiong et al., 2024), LC vs RAG superiority in long-form medical QA remains an open question. We make the following contributions:

- We are the first to assess how LLMs of different underlying capabilities and sizes perform on long-context medical QA given longitudinal EHR patient notes, covering three task formulations: 1) MCQA , 2) Extractive and 3) Open-ended generative QA (§3.1) and a variety of LC and RAG settings, as depicted in Fig. 1.
- We perform comprehensive experiments evaluating different note inclusion strategies in the input context, including note memorization (§4.2).
- We carefully design a hybrid RAG pipeline examining its performance over Full-context (FC) across task formulations and context sizes, showing its superiority across tasks (§5.1,§5.2).
- Our experiments and multi-faceted evaluation methodology allow for a direct comparison between long-form reasoning when the answer is part of a single versus multiple documents (§5.1).
- Our quantitative and qualitative analyses uncover some of the LLM and metrics challenges in reasoning over long EHRs, while demonstrating such cases where RAG excels (§5.2,§5.3).

2 Related Work

2.1 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 passage retrieval) (Zhang et al., 2024b; Chen et al., 2025b). Despite positional embeddingtechniques like RoPE (Su et al., 2024b), YaRN (Peng et al., 2023) and Position Interpolation (Chen et al., 2023a) allowing context extrapolation, QA task performance is low even for 32K contexts (Chen et al., 2025b). Furthermore, the performance gap between open-source and longer-context close-source models (Li et al., 2024a) poses questions around the potential avenues for patient-centric long-dependency data where privacy matters.

Assessment of long-context LLM performance (Dong et al., 2023; Bai et al., 2023; Liu et al., 2023; Zhang et al., 2024b; Hsieh et al., 2024), shows that apart from performance drops with increasing token length, LLMs are particularly challenged by long-range dependencies (Li et al., 2024a) that potentially require reasoning. Fan et al. (2024) show notable LLM performance drop in medical QA tasks with contexts up to 200K tokens, with open-source LLMs struggling to produce output given larger contexts. Although long and longitudinal context is prominent in real-world patient and medical data, there is little work investigating LLM capabilities in such settings.

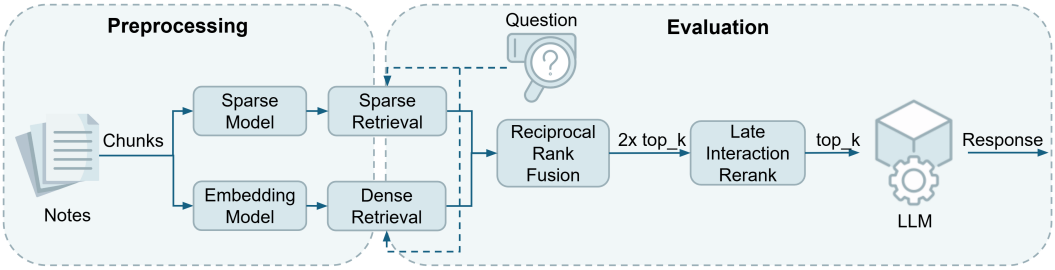

Figure 2: RAG Pipeline

2.2 Retrieval-augmented Generation (RAG)

RAG has been leveraged in a prompt-based plug-in manner to address black-box long-context challenges (Yu et al., 2023). Current retrieval strategies are primarily: chunk-based (Izacard et al., 2021), index-based (Liu, 2022) and summarization-based (Sarthi et al., 2024). LLM-based large-scale retrievers for chunk-based approaches like BGE M3 (Chen et al., 2024a) give notable performance improvement. Furthermore, approaches like Self-RAG (Asai et al., 2023) outperform ChatGPT in open-domain QA through on-demand retrieval followed by generation and self-reflection.

Biomedical retrievers showed benefits in domain retrieval, with the MedCPT dense semantic retriever and re-ranker (Jin et al., 2023), achieving top performance on 6 biomedical tasks surpassing LLM-based counterparts. The BMRETRIEVER (Xu et al., 2024b) further surpasses MedCPT's performance on most downstream tasks. More recently, methods like Med-RAG (Zhao et al., 2025) showed better diagnostics over EHRs, through enhancement by knowledge graph-elicited reasoning, while Self-BioRAG (Jeong et al., 2024) employs MedCPT and generates through self-reflection using a domain-specific LLM. Yet Fan et al. (2025) underscore the medical retrieval challenges despite strong in-domain retrieval capabilities across large corpora, with models struggling with specialized medical content such as EHRs. Myers et al. (2025) further demonstrated variability in retrieval across EHR datasets. Therefore a thorough examination of RAG vs LC over realistic long-context patient-oriented QA is needed.

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** (see §4.2).

3.1 Dataset Selection Criteria

The following criteria were used for benchmark selection in our study (full comparison in Table 10):

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 clinical settings. Our evaluation goes beyond MCQA to 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., 2024) or synthetic data (Adams et al., 2024).

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.

3.2 RAG Approach

Retrievers are distinguished into *sparse* and *dense*. Sparse retrievers such as BM25 (Robertson et al., 2009) and Splade (Formal et al., 2021) use overlapping terms to match queries with snippets, while dense retrievers encode them into embeddings and match them based on semantic similarity. Recently,

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. Additionally, non-hybrid experiments on EHR subsets of varying granularity (Fan et al., 2025) show no clear winner between sparse and dense retrievers. 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). Based on the abov, we leverage a hybrid approach followed by re-ranking, presented in Figure 2. 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 [1]:

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

Dense retriever: we use the same Encoder, E, for the query and the chunks and capture semantic similarity of pairs using the cosine similarity:

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

Reciprocal Rank Fusion (RRF): combines the top-k chunks from each retriever accounting for each document's rank, 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 is semantically supported by at least one document token, to produce the final 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 temporally rather than by retrieval score due to its superior long-context performance (Yu et al., 2024).

3.3 Evaluation Methodology

Evaluating medical QA systems has traditionally relied on automatic 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. Towards 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 (e.g., MEDIC (Kanithi et al., 2024)) 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 expert verified/generated enabling clinically-centric evaluation:

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

BERTScore (Zhang et al., 2020), 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.

Accuracy reported on the MCQA task.

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

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

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.

overview below, also summarized in Table 1:

CliniQG4QA (Yue et al., 2020): An extractive span benchmark based on MIMIC-III. While most of its pairs are machine-generated (MG), the 1,287 QA test pairs are either generated or verified by three clinical experts using the MG questions for reference. We only use the expert-annotated set.

RadQA (Soni et al., 2022): An extractive dataset comprising 3,074 physician-crafted questions and 6,148 answer spans from the Findings and Impressions sections of RR in MIMIC-III. Many questions correspond to multiple sentences and require different types of reasoning. Annotation is carried out by two human annotators. After filtering out for *Unanswerable* pairs we obtain 3,509 pairs.

EHR-DS-QA (Kotschenreuther, 2024): A synthetic QA corpus of 156,599 pairs generated by two LLMs and guided by predefined prompt templates on DS and CN from MIMIC-IV. A subset of 506 pairs were physician-verified of which we retain 478, marked as correct.

EHRNoteQA (Kweon et al., 2024): Built on MIMIC-IV, it contains 962 QA pairs authored by GPT-4 and iteratively refined by three clinicians. Questions span a diverse set of topics, each corresponding to multiple DS (∼2.3 per patient), while supporting both open-ended and MCQA formats.

4.2 Context Formulation

While most of the datasets are based on a single note/report (CliniQG4QA, RadQA, EHR-DS-QA) or up to three DS (EHRNoteQA), there is a large amount of *longitudinal patient supplementary* content that remains unused. Aiming for a realistic clinical setting where there is no a-priori knowledge of the most relevant note, we leverage the MIMIC resources to augment the context with each patient's longitudinal notes[2], resulting in long context 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 Relevant**: Exclusion of relevant notes - assessing usefulness of supplementary material.
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) as marked in each dataset - assessing model's ability to leverage relevant context information.

For the *Include All* setting we report context-length-based performance across four token bins: *Short context*: 0–8K, *Medium context*: 8–16K, *Large context*: 16–32K, *Extended context*: 32–128K.

4.3 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., 2024b): 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 stability and agreement (see Appendix D.2).

Retrieval Settings: Based on biomedical QA retrieval research that demonstrates the superiority of combining sparse and dense retrievers (see §3.2), we designed a hybrid retrieval approach to properly handle the semantic, lexical, and query complexity. The selected model components are enlisted below:

- *Dense Retrieval* using the open embedding model Qwen3-Embedding-8B (Zhang et al., 2025).
- *Sparse Retrieval* using BM25 selected through two sparse retriever ablation (see Table 9).
- *Late Interaction Reranking* with Reason-ModernColBERT, a ColBERT (Khattab and

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

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	11.3	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	23.8	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	**29.9**	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	83.8
QwQ 32B		**75.52**	47.14	81.02	89.25	57.68	60.54	78.9	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. Red highlights the global best across all Settings per Metric. *LLM* corresponds to LLM Correctness, *NLI* to NLI Entailment and *F1* to BioBERT F1

Zaharia, 2020) model finetuned on the ReasonIR dataset (Shao et al., 2025) demonstrating high performance on reasoning-intensive retrieval benchmark (Su et al., 2024a).

We explored two chunk inclusion strategies: i) direct chunk inclusion (RAG), ii) hierarchical parent note inclusion (RAG HIR).

Experimental Setup: LLM inference and prompt formulation specifics are in Appendices A and C.3.

5 Results and Discussion

5.1 Main Results

Performance for the different inclusion settings across datasets and models is shown in Table 2.

Model Size, Tasks and Inclusion Settings: Consistent with general findings, larger models perform better across all tasks (Table 2). Qwen2.5:32B and QwQ:32B generally outperformed the 7B parameter models across most metrics and tasks. The performance gap was particularly pronounced in tasks requiring reasoning or information synthesis from multiple sources, namely EHRNoteQA and RadQA. However, smaller models performed better in settings excluding relevant information or completely removing the context. This suggests that they are more prone to memorization. Additionally, unlike the rest of the tasks, 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 'Include Related' setting outperforms 'Include All' on most datasets and metrics except for MC. Despite that, the 'Exclude Related' setting outperforms the 'Exclude All' by a margin across the board showing that the usefulness of the supplementary patient material and suggesting that while the 'Include All' setting includes useful additional information, LLMs struggle to process long context and identify the most relevant information.

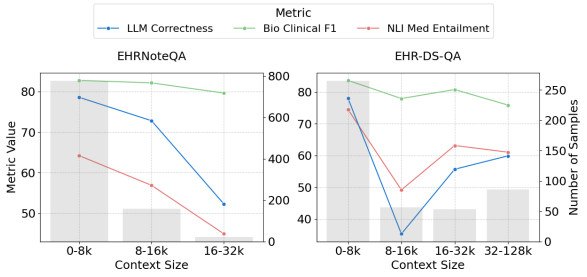

Figure 3: Metric performance over context size on 'Include All' for Qwen2.5:32B-Instruct.

Reasoning and Fine-tuning: The impact of reasoning-focused models and medical fine-tuning is mixed. *The QwQ:32B general reasoning model* showed improved results on Table 2 over other models overall on RadQA, EHRNoteQA - the reasoning datasets. This suggests that reasoning models can offer an advantage in tasks that demand complex inference across multiple sources of information, but this benefit may be more apparent in larger model sizes where reasoning capabilities can be effectively leveraged without compromis-

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
	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	70.8	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
	RAG 5	59.54	40.33	78.82	75.67	54.17	56.11	**79.82**
	RAG 10	64.03	45.02	79.05	73.4	57.11	**62.76**	79.6
	RAG 15	58.07	52.08	79.20	75.6	**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
	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. Red highlights the global best across all models per Metric.

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 and red highlights the global best per Metric.

ing other essential skills. By contrast, *the medically fine-tuned reasoning model, HuatuoGPT-o1-Qwen2.5:7B*, did not show any improvement over its standard instruction-tuned base model Qwen2.5:7B-instruct, in line with studies revealing that biomedical LLMs often lead to reduced performance (Dorfner et al., 2024; Dada et al., 2025).

Context Size: Fig. 3 shows results for Qwen2.5:32B-Instruct over different context sizes. Generally, performance decreases across metrics with increased context size on open-ended QA particularly when multi-note reasoning is required (EHRNoteQA), showing that even large LLMs are challenged by long context, with similar trends observed across models (see Appendix, Fig. 8). Worth noting that for context range (8-16K) EHR-DS-QA exhibits a dip in performance across all metrics and models. Such dip is likely attributed to data noise, based on manually observed inconsistencies and ambiguous gold answers in some pairs especially in EHR-DS-QA. correlating with strange dips in performance at mid-range context windows. Fan et al. (2024) and Ma et al. (2025) observed similar dips.

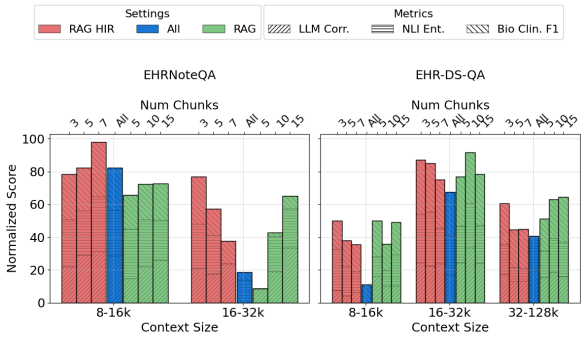

Figure 4: RAG performance across settings with Qwen2.5:32B-Instruct (min-max normalized).

RAG Performance: Table 3 shows the two evaluated chunk settings on the open-ended generative datasets[3]. *RAG demonstrated clear performance improvement for the single-note generative task of EHR-DS-QA*. In this scenario, the chunk-in-context strategy consistently provided the best results. *For the more complex, multi-note reasoning task of EHRNoteQA, the hierarchical RAG strategy (RAG HIR) was the best* with few exceptions. The results also show similar patterns on MC indicating clear performance improvement for all models on the RAG HIR 7 setting. Our findings are consistent across medium and larger (8K+) context sizes (Fig. 4). Worth noting that using RAG, the medically fine-tuned reasoning model HuatuoGPT-o1-Qwen2.5:7B scored the best results on multiple metrics across the open-ended datasets, showing that RAG could benefit even more so 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.

5.2 Quantitative Analysis: RAG vs LC

In our analysis we gathered challenging cases of disagreement between metrics (LLM Correctness, NLI Entailment, and Bio-F1) across EHRNoteQA and EHR-DS-QA. Metrics were first z-score standardized and samples were gathered for cases were metrics belonged to the top 60th percentile while

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

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: Breakdown of sampled favored by RAG vs FC across four categories in open-ended QA.

another fell within the bottom 40th percentile. The 41 examples gathered were then manually examined in order to categorize them across QA types and determine which answer between RAG and FC is more reliable. The rubric is provided below:

- **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.

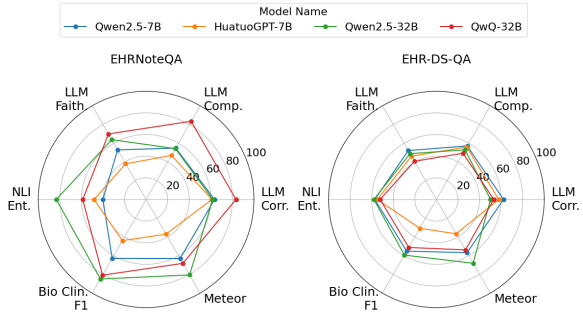

Figure 5: Model performance across metrics for open-ended QA, normalized with min-max per metric.

In Table 5, RAG shows higher performance across the SEE, CNS and ANI categories, while FC leads in the TLR one, demonstrating that while RAG effectively preserves the relevant information for challenging cases, it falls behind when temporal reasoning is required. A further qualitative analysis providing insights on cases where RAG or FC perform better is provided in Appendix G[4].

5.3 Metrics and Insights

Metric Comparisons: Fig. 5 provides insights into the relation between models, metrics, and datasets while focusing on the open-ended generative tasks for the 'Include All' and 'Include Related' settings combined. *Instruct models lead on the single-note non-reasoning task (EHR-DS-QA), while larger models lead on the multi-note reasoning one (EHRNoteQA)*. Instruct models are better based on semantic and NLI metrics, while reasoning ones score better on LLM-as-a-judge metrics[5].

Metric Insights: We performed a qualitative analysis to investigate the reasons behind metrics' disagreements on our open-ended datasets (see Appendix E.2). The 233 disagreements in EHRNoteQA and 47 in EHR-DS-QA are mainly due to:

- Correct/Complete but Unfaithful answers with addition of unsupported details in both datasets but more pronounced in EHR-DS-QA.
- Correct answer but low surface overlap (lower Meteor/Bert) in both datasets.
- Correct answer but low NLI entailment occurs due to negation/clinical phrasing in both datasets.
- Several short, risk/negation items where NLI flags contradiction despite clinically aligned answers in EHR-DS-QA.

These point to the need for thinking about more reasoning appropriate metrics.

6 Conclusion

We studied long-context, patient-centric clinical QA across EHR-grounded datasets, contrasting Full-Context (FC) prompting with Retrieval-Augmented generation (RAG). Our work shows that highly-relevant context often outperforms feeding all notes, likely due to LLMs struggling with long context. Hybrid RAG pipelines with reranking overperform FC, especially for specific factual queries and multi-note synthesis. Larger models generally help especially for reasoning datasets, but performance remains sensitive to context length and task formulation, while medically fine-tuned models fall behind. Future directions include: (i) scaling beyond single-note or single-source assumptions toward richer multi-note, multi-visit reasoning; (ii) advancing temporal and causal reasoning over longitudinal records, with explicit timeline grounding; (iii) strengthening evaluation via clinically faithful, judge-robust rubrics and cross-metric reliability analyses.

[4] manually analyzed failure cases in Appendix F.

[5] We include all-context and long-context correlations between metrics in Appendix E.1

Limitations

Our work focuses on MIMIC-derived datasets, which represents English-only data in 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 the 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 complexities of modeling clinical notes, real-world records comprise heterogeneous data, including data from other sources and in other modalities. Our work poses limitations in the assessment of long-context by focusing only on the textual modality, something we aim to address in future work.

Despite the in depth study of literature and our initial ablations in selecting a competitive RAG methodology, our work does not exhaustively examine the full potential of RAG under different settings, i.e. different retrievers and chunk sizes, in long-form medical QA. 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. Our analysis has also pointed to the limitations of current evaluation metrics, both in terms of assessing reasoning as well as lack of transparency in LLM as a judge metrics and how faithfulness can be truly assessed currently when additional information, not present in the benchmark, is added.

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.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, An drey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh,

Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Lisa C. Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Loser, Hugo J. W. L. Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Kyrill Bressem. 2024. Longhealth: A question answering benchmark with long clinical documents. *ArXiv*, abs/2401.14490.

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*.

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.

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. *arXiv preprint arXiv:2502.06666*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *ArXiv*, abs/2310.11511.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

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. 2024. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *ArXiv*, abs/2412.15204.

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.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, T. W. Hennigan, Saffron Huang, Lorenzo Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and L. Sifre. 2021. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*.

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.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Annual Meeting of the Association for Computational Linguistics*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024b. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Pei Chen, Hongye Jin, Cheng-Che Lee, Rulin Shao, Jingfeng Yang, Mingyu Zhao, Zhaoyu Zhang, Qin Lu, Kaiwen Men, Ning Xie, et al. 2025b. 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.

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, Jens Kleesiek, et al. 2025. Does biomedical training lead to better medical performance? In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 46–59.

Pritam Deka, Anna Jurek-Loughrey, et al. 2023. Multiple evidence combination for fact-checking of health-related information. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 237–247.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv preprint arXiv:2309.13345*.

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*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aur'elien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cris tian Cantón Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab A. AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriele Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guanglong Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Ju-Qing Jia, Kalyan Vasuden Alwala, K. Upasani, Kate Plawiak, Keqian Li, Ken-591 neth Heafield, Kevin R. Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuen ley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melissa Hall Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Niko lay Bashlykov, Nikolay Bogoychev, Niladri S. Chatterji, Olivier Duchenne, Onur cCelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasić, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Ro main Sauvestre, Ron nie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sa hana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Chandra Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Vir ginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit ney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yiqian Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zhengxu Yan, Zhengxing Chen, Zoe Papakipos, Aaditya K. Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adi Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Po-Yao (Bernie) Huang, Beth Loyd, Beto de Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina

Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Shang-Wen Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzm'an, Frank J. Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory G. Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kaixing(Kai) Wu, U KamHou, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pe dro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollár, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sung-Bae Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Andrei Poenaru, Vlad T. Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xia Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Yongqi Fan, Hongli Sun, Kui Xue, Xiaofan Zhang, Shaoting Zhang, and Tong Ruan. 2024. Medodyssey: A medical domain benchmark for long context evaluation up to 200k tokens. *arXiv preprint arXiv:2406.15019*.

Yongqi Fan, Nan Wang, Kui Xue, Jingping Liu, and Tong Ruan. 2025. 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.

S. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo Pontes Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan H. Steinberg, Ashwin Nayak, Birju S. Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott J. Adams, Oluseyi Fayanju, Shreya Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher D. Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma Brunskill, Jason Alan Fries, and Nigam H. Shah. 2023. Medalign: A clinician-generated dataset for instruction following with electronic medical records. In *AAAI Conference on Artificial Intelligence*.

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*, pages 2288–2292.

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*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin,

and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

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(Supplement_1):i119–i129.

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 W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

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 Johnson, Lucia Bulgarelli, Tom Pollard, Steve Horng, Leo A. Celi, and Roger Mark. 2023. MIMIC-IV-Note: deidentified free-text clinical notes (version 2.2). *PhysioNet*.

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

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*, pages 39–48.

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. *arXiv preprint arXiv:2405.01535*.

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, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. In *Neural Information Processing Systems*.

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. IEEE.

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. *arXiv preprint arXiv:2402.14848*.

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. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

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.

Xinze Li, Yixin Cao, Yubo Ma, and Aixin Sun. 2024b. Long context vs. rag for llms: An evaluation and revisits. *arXiv preprint arXiv:2501.01880*.

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

Jerry Liu. 2022. Llamaindex. *CoRR*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

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.

Zizhan Ma, Wenxuan Wang, Guo Yu, Yiu-Fai Cheung, Meidan Ding, Jie Liu, Wenting Chen, and Linlin Shen. 2025. Beyond the leaderboard: Rethinking medical benchmarks for large language models. *arXiv preprint arXiv:2508.04325*.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A Rossi, Seunghyun Yoon, and Hinrich Schütze. 2025. Nolima: Long-context evaluation beyond literal matching. *arXiv preprint arXiv:2502.05167*.

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, et al. 2025. Overview of bioasq 2025: The thirteenth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 173–198. Springer.

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 *Clinical Natural Language Processing Workshop*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Anusri Pampari, Preethi Raghavan, Jennifer J. Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. In *Conference on Empirical Methods in Natural Language Processing*.

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.

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, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. *ArXiv*, abs/2401.18059.

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, and Luke Zettlemoyer. 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.

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. *arXiv preprint arXiv:2401.16240*.

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.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik, Danqi Chen, and Tao Yu. 2024a. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024b. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Simon Šuster and Walter Daelemans. 2018. Clicr: a dataset of clinical case reports for machine reading comprehension. *arXiv preprint arXiv:1803.09720*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark B. Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *ArXiv*, abs/2311.10537.

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.

Shubham Vatsal and Ayush Singh. 2024. Can gpt redefine medical understanding? evaluating gpt on biomedical machine reading comprehension. *ArXiv*, abs/2405.18682.

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.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj digital medicine*, 6(1):135.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

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.

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 D Wang, Joyce C Ho, Chao Zhang, and Carl Yang. 2024b. Bmretriever: Tuning large language models as better biomedical text retrievers. *arXiv preprint arXiv:2404.18443*.

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 M. Pournejatian, Hoo-Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin B. Compas, Cheryl Martin, Anthony B Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria P. Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *NPJ Digital Medicine*, 5.

Tan Yu, Anbang Xu, and Rama Akkiraju. 2024. In defense of rag in the era of long-context language models, 2024. *URL https://arxiv. org/abs/2409.01666*.

W. Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023. Improving language models via plug-and-play retrieval feedback. *ArXiv*, abs/2305.14002.

Xiang Yue, Xinliang Frederick Zhang, Ziyu Yao, Simon M. Lin, and Huan Sun. 2020. Cliniqg4qa: Generating diverse questions for domain adaptation of clinical question answering. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 580–587.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. 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.

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*, pages 4442–4457.

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.

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
	k_1	TF saturation
	b	length norm.
	avgdl	avg. doc length

Table 6: Hyperparameter Configuration

C Datasets

C.1 Context Data Distributions

Token-length distributions per dataset are shown in Figures 6 and 7.

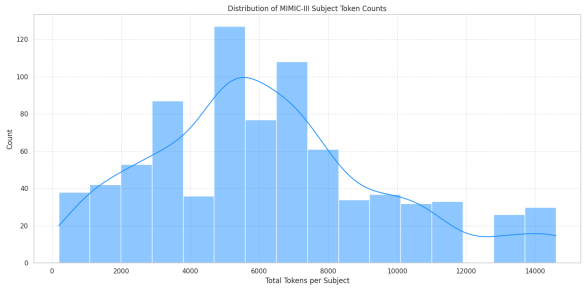

Figure 6: MIMIC-III Context Distributions

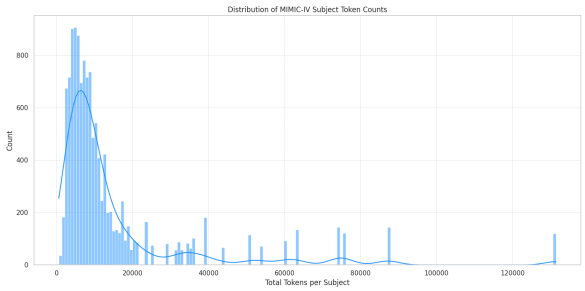

Figure 7: MIMIC-IV Context Distributions

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.
- 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.

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

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 8: 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 9: Comparison of Different Sparse Retrievers

C.3 Prompt Formulations

We used a simple one 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.

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.4 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 10 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 (§4.2), enabling a realistic comparison of full-context and RAG pipelines. We provide citations of each considered dataset from Table 10 in Table 11.

D Model Comparisons

D.1 Retrievers

On Table 9 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 LLM-as-a-judge

In Table 7 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 [6] has far less such disagreement compared to more specialized judges. Table 8 also provides direct comparisons of LLM-as-a-judge model, demonstrating cases where LLM Correctness between such pairs is more than 0.5 and therefore high. Selene-

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

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 10: Biomedical Datasets Comparison. References for each dataset are in Table 11.

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

Table 11: Corresponding Citations for Datasets considered in Dataset Selection of Table 10

8B [7] and Prometheus-8x7B-v2.0 [8] comparisons show that these two models are consistently more in agreement between them while Qwen2.5-32B-Instruct shows more independent judging abilities.

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

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

D.3 Context Size Performance

Figure 8 presents performance for on open-ended QA across metrics and models.

E Metrics

E.1 Correlations

We examine correlations between evaluation metrics to identify metric independence. On EHRNoteQA and EHR-DS-QA across all non-exclude settings (Figure 9) and long-context high LLM Correctness non-exclude settings (Figure 10), LLM evaluation metrics exhibited strong inter-correlations and METEOR and BERTScore were highly correlated. Although the above correlations remain high, they are less pronounced for the long-context high LLM Correctness setting of Figure 10. In general all metrics are less correlated for the longer context setting.

E.2 Analysis

Table 12 showcases a qualitative analysis on what metrics are capturing vs missing based on analyzing disagreement data.

F 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

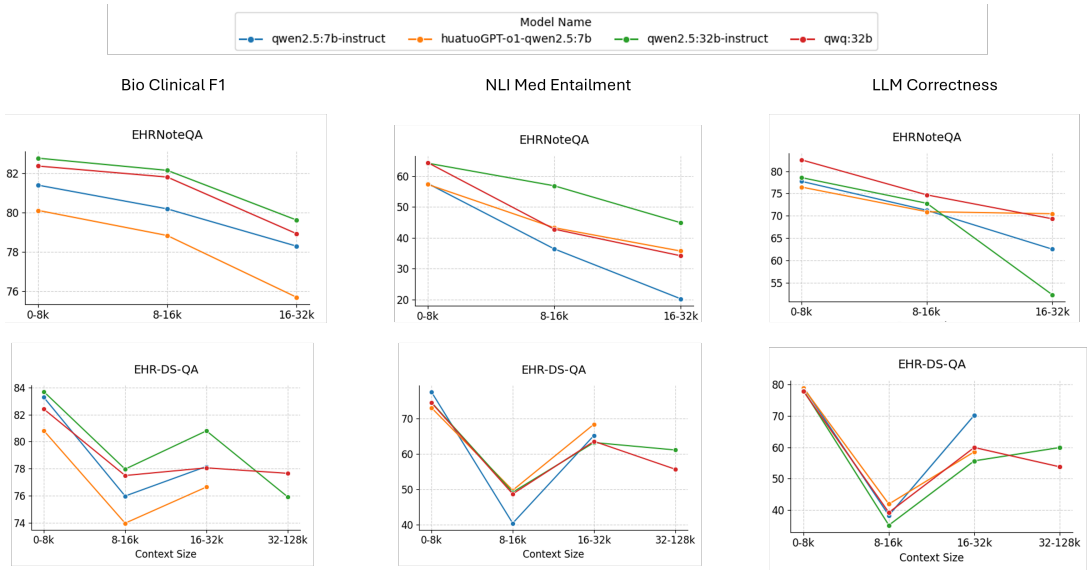

Figure 8: Model performance over context size.

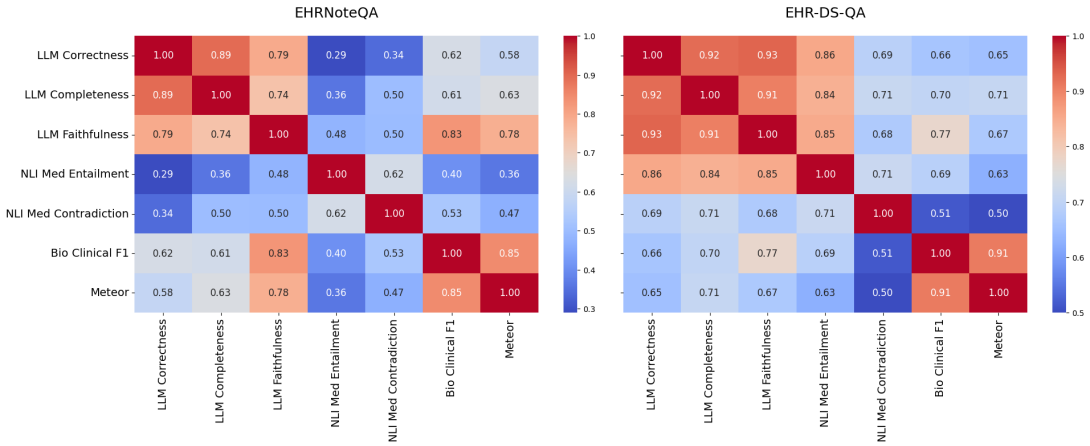

Figure 9: Metric correlations across all non-exclude settings.

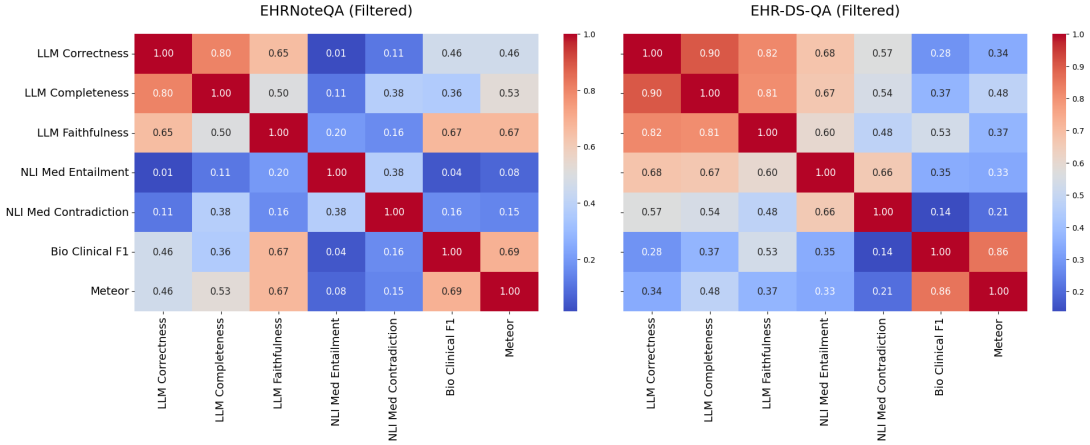

Figure 10: Metric correlations across all non-exclude settings filtered by LLM Correctness $>$ 50 and Context $>=$ 16K.

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 12: What each metric captures vs. misses, with typical disagreement patterns and representative examples from EHRNoteQA and EHRDSQA.

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 → increased ostomy output → severe dehydration (prerenal) → 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.

G RAG vs FC: Detailed Analysis

We summarize our insights about which RAG or FC setting tends to be advantageous across datasets and query types in Table 13. 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, such as summaries, 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 information scattered 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 with 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.

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	FC is better 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 13: RAG vs FC Detailed Analysis.