LAMAR at ArchEHR-QA 2025: Clinically Aligned LLM-Generated Few-Shot Learning for EHR-Grounded Patient Question Answering

Seksan Yoadsanit*[1,2], Nopporn Lekuthai*[1,2], Watcharitpol Sermsrisuwan[1,2], Titipat Achakulvisut[1]

[1] Department of Biomedical Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom, Thailand

[2] Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

Correspondence: titipat.ach@mahidol.ac.th

Abstract

This paper presents an approach to answering patient-specific medical questions using electronic health record (EHR) grounding with ArchEHR-QA 2025 datasets. We address medical question answering as an alignment problem, focusing on generating responses factually consistent with patient-specific clinical notes through in-context learning techniques. We show that LLM-generated responses, used as few-shot examples with GPT-4.1 and Gemini-2.5-Pro, significantly outperform baseline approaches (overall score = 49.1), achieving strict precision, recall, and F1-micro scores of 60.6, 53.6, and 56.9, respectively, on the ArchEHR-QA 2025 test leaderboard. It achieves textual similarity between answers and essential evidence using BLEU, ROUGE, SARI, BERTScore, AlignScore, and MEDCON scores of 6.0, 32.1, 65.8, 36.4, 64.3, and 43.6, respectively. Our findings highlight the effectiveness of combining EHR grounding with few-shot examples for personalized medical question answering, establishing a promising approach for developing accurate and personalized medical question answering systems. We release our code at https://github.com/biodatlab/archehr-qa-lamar.

1 Introduction

Large language models (LLMs) have significantly influenced medical question-answering systems by generating clinically relevant content grounded in electronic health records (EHRs) for more personalized and context-aware patient care (Yang et al., 2022). Clinical-related questions are among the most frequently asked topics, reflecting the public's natural curiosity about their health literacy and the rising healthcare costs in many countries, which drive individuals to seek alternative sources of information (Savery et al., 2020). Despite some hallucinations, recent frontier models typically maintain reasonable factual accuracy. We theorized that aligning with human expectations on answering style, citation practices, and information structuring is the main challenge. Thus, we formulate our approach to align the model response to human expectation with the limited data provided in this shared task.

While fine-tuning LLMs on medical records or textbooks can improve alignment, it demands extensive datasets, limiting scalability (Singhal et al., 2023). Few-shot learning offers a promising alternative by guiding models with representative examples that demonstrate task-specific reasoning patterns without requiring fine-tuning, though designing optimal examples remains challenging (Brown et al., 2020). Similarly, Retrieval-augmented generation (RAG) enables LLMs to access external knowledge sources such as structured medical databases and clinical literature, providing accurate, up-to-date answers by incorporating the medical knowledge without retraining (Alkhalaf et al., 2024; Lewis et al., 2020). However, questions remain about how effectively retrieved information is integrated and grounded in the model's final output, particularly in clinical contexts where alignment with human preferences is crucial.

In this paper, we present an approach to answering patient-specific medical questions using electronic health record (EHR) grounding with the ArchEHR-QA 2025 dataset (Soni and Demner-Fushman, 2025b). We address medical question answering as an alignment problem, focusing on generating responses factually consistent with patient-specific clinical notes through in-context learning techniques. Our system leverages LLM-generated responses as few-shot examples with GPT-4.1 and Gemini-2.5-Pro, achieving strict precision, recall, and F1-micro scores of 60.6, 53.6, and 56.9, respectively, on the test leaderboard.

*Equal Contribution

2 Related work

Large Language Models (LLMs) have demonstrated significant potential across diverse medical question-answering applications. Initial research focused on general medical knowledge retrieval (Shi et al., 2024), while subsequent work has expanded into specialized domains including USMLE-style multiple-choice questions (Lucas et al., 2024), clinical decision support (Benary et al., 2023), medical exam preparation (Artsi et al., 2024), and patient-facing information systems (Goodwin et al., 2022). Despite these advances, LLMs continue to face challenges. Hallucinations remain a key concern in medical settings (Agarwal et al., 2024), and newer models have made progress in reducing them (Kim et al., 2025). However, real-world EHRs introduce an even bigger hurdle: clinical data are often messy, incomplete, and inconsistent (Holmes et al., 2021). Issues such as outdated knowledge and inconsistent reasoning also persist and demand ongoing attention (Ji et al., 2023).

In-context learning (ICL) provides an efficient alternative to model fine-tuning, enabling LLMs to learn from demonstrations embedded directly in prompts without requiring parameter adjustments. Dong et al. (2022) demonstrate that ICL leverages pre-trained capabilities to recognize task patterns from limited examples, reducing dependency on supervised datasets (Dong et al., 2024). Few-shot prompting, popularized by Brown et al. (2020) with GPT-3, showed that LLMs can achieve competitive performance across diverse tasks, including medical question answering, by conditioning on carefully selected examples. This approach significantly reduces barriers to adapting LLMs for specialized applications like clinical reasoning without requiring domain-specific retraining or extensive annotated data (Brown et al., 2020).

Alkhalaf et al. demonstrated that combining generative AI with Retrieval-Augmented Generation (RAG) significantly improves clinical information extraction from EHRs, achieving 99.25% accuracy using LLaMA 2 13B with zero-shot prompting (Alkhalaf et al., 2024). Beyond methodology, RAG component quality is critical for performance, as highlighted by research using the MEDRAG toolkit across 41 configurations with varying models, retrievers, and knowledge corpora. This comprehensive analysis revealed that properly implemented RAG systems can boost accuracy by up to 18% compared to chain-of-thought prompting across multiple medical question-answering tasks, emphasizing the importance of appropriate language model selection, retrieval strategy, and knowledge corpus construction (Xiong et al., 2024).

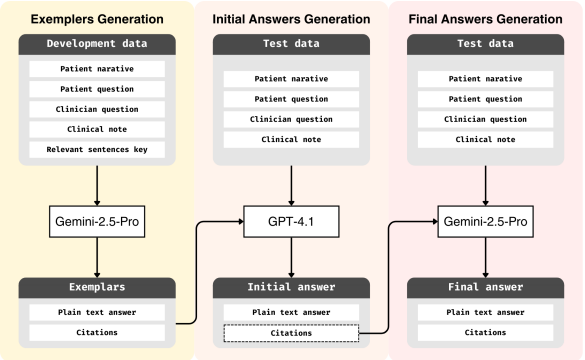

Figure 1: The multistage few-shot prompting pipeline of our system submitted to ArchEHR-QA 2025.

3 Datasets

We utilize the ArchEHR-QA 2025 dataset (Soni and Demner-Fushman, 2025a) , which comprises 20 development cases and 100 testing question-note pairs. The dataset includes a patient question, a clinician question, and a clinical note. The development set features ground-truth annotations for evidence sentences, while the test set requires natural language answers accompanied by cited sentence numbers. The patient's question is inspired by real patient inquiries. Clinical note excerpts are derived from the MIMIC-III database (Johnson et al., 2016). Answers consist of sentences referenced by the ID from the clinical note.

4 Methodology

We viewed the problem as an alignment issue. We aimed to generate an answer that was correctly cited and factually aligned with the clinical note. To help generate aligned answers, we explored zero-shot, few-shot prompting, and retrieval augmented generation (RAG) with external sources including MedlinePlus and Merck Manual.

4.1 Baseline

We applied zero-shot and chain-of-thought (CoT) prompting conditions as our baselines. We select non-thinking models, including GPT-4.1, Gemini-2.0-Flash, or Claude-3.7-Sonnet (non-thinking), due to their significant computational and financial overhead. Each model was prompted to reason

step by step before generating a final answer, providing a reference for measuring the impact of few-shot examples and retrieval-augmented generation (RAG).

4.2 In-Context Learning through Few-shot Prompting

We explored several few-shot prompting strategies, including:

- **Basic Few-shot.** We selected two examples from the ArchEHR website as few-shot.
- **LLM-Generated Exemplars as Few-shot.** Since the relevant sentence labeling can only be found in the development dataset, we used Gemini-2.5-Pro to generate answers from the development set. These answers, paired with their corresponding clinical notes and questions, were used as few-shot examples in subsequent prompts.
- **LLM-Generated Exemplars with Reasoning.** We want to see if examples with reasoning can help improve the answer. Here, we used Gemini-2.5-Pro to generate both reasoning steps and final answers. These reasoning-annotated examples were included in prompts to simulate clinical thinking.

4.3 Retrieval-Augmented Generation (RAG)

We tested external context enhancement using 10,232 MedlinePlus (National Library of Medicine (US), 2025) and 2,927 Merck Manual articles (Merck & Co., 2025). This experiment aimed to determine whether external medical knowledge could improve answer accuracy over a few-shot exemplar. Articles were embedded using MedCPT (Jin et al., 2023) and indexed for retrieval. We compared 3 retrieval approaches:

- **Full-Text Clinical Articles.** We retrieved the complete texts of relevant clinical publications identified by our RAG pipeline. We input the entire article to provide the model with comprehensive contextual information for answer generation.
- **Concise Article Summaries.** We prompted Gemini-2.5-Pro to distill each full-text article into a one-paragraph summary to reduce prompt length and boost information density.
- **Synthetic Clinical Cases.** We prompted Gemini-2.5-Pro with a few-shot examples from the ArchEHR page to transform and format the retrieved articles into a realistic patient scenario featuring patient narrative, patient question, clinical question, clinical notes, and answer to mimic the ArchEHR dataset.

5 Evaluation

Each answer includes sentences and their references to the clinical note. Generated sentences are evaluated on factuality and relevance. Factuality compares cited evidence to ground truth using precision, recall, and F1 scores, with both strict (essential sentences only) and lenient (essential and supplementary sentences). Relevance measures textual similarity between answers and essential evidence using BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), SARI (Xu et al., 2016), BERTScore (Zhang et al., 2019), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023). The final score averages normalized Strict Citation F1 and composite Relevance.

6 Experimental Setup

We use zero-shot and chain-of-thought prompting with GPT-4.1, Gemini-2.0 Flash, and Claude-3.7-Sonnet as our baseline. In few-shot prompting, we use Gemini-2.5-Pro to generate 19 exemplar answers for each item in the development set, excluding the item itself. Generated answers are sent to Gemini-2.5-Pro to trim and summarize answers to a 75-word limit. We set up top-k=5 in retrieval for all RAG experiments.

7 Results and discussions

We evaluated various prompting and retrieval strategies on the development dataset to assess their impact on citation accuracy, factuality, and relevance across multiple language models. These experiments informed our final multi-stage pipeline design for the test dataset. The following sections present key results and their implications for our system.

7.1 Development dataset observation

7.1.1 Zero-shot Baselines

On the development set, GPT-4.1, Gemini-2.0-Flash, and Claude-3.7-Sonnet achieved overall zero-shot scores of 47.9, 45.0, and 48.1, respectively, with Claude-3.7-Sonnet performing best (Ta-

Approach	Model	Development dataset																			
		Overall	Factual.	Relev.	SP_μ	SR_μ	SF_μ	LP_μ	LR_μ	LF_μ	SP_M	SR_M	SF_M	LP_M	LR_M	LF_M	BLEU	ROUGE-L	SARI	BERTScore	AlignScore	MEDCON
Baseline	GPT-4.1	47.9	53.8	42.1	54.0	53.6	53.8	68.6	49.7	57.7	60.4	64.1	57.1	74.2	59.4	62.2	6.9	34.1	69.5	37.3	61.8	42.7
	Gemini-2.0-Flash	45.0	49.2	40.8	55.5	44.2	49.2	70.0	40.7	51.5	60.2	53.7	53.5	73.0	50.7	56.2	6.3	30.6	65.4	35.9	65.4	41.0
	Claude-3.7-Sonnet	48.1	57.1	39.2	51.1	64.5	57.1	65.5	60.3	62.8	52.6	72.5	57.5	66.9	68.0	64.1	5.7	31.6	64.2	35.6	56.6	41.2
Few-shot: basic	GPT-4.1	48.6	57.8	39.4	**60.8**	55.1	57.8	72.8	48.1	58.0	**65.5**	64.5	59.2	**79.0**	59.6	62.5	5.5	31.4	65.5	35.6	57.1	41.3
	Gemini-2.0-Flash	47.5	56.3	38.8	54.0	58.7	56.3	66.0	52.4	58.4	56.4	66.4	56.2	70.5	62.3	61.3	5.2	30.2	65.5	33.8	56.3	41.9
	Claude-3.7-Sonnet	48.1	55.6	40.6	49.7	63.0	55.6	62.3	57.7	59.9	55.1	73.9	57.4	67.8	70.1	63.5	7.0	32.8	66.3	35.2	60.7	41.5
Few-shot: LLM-generated exemplars	GPT-4.1	**51.5**	**61.1**	41.8	56.4	66.7	**61.1**	71.8	61.9	**66.5**	60.3	**77.2**	**63.6**	76.9	74.0	**71.4**	6.3	32.7	66.5	36.4	64.0	44.9
	Gemini-2.0-Flash	50.5	59.6	41.5	54.9	65.2	59.6	65.9	57.1	61.2	56.9	75.9	59.8	68.8	69.5	63.6	7.5	32.1	66.7	36.2	60.9	45.6
	Claude-3.7-Sonnet	49.6	58.4	40.7	51.1	**68.1**	58.4	65.2	63.5	64.3	53.4	76.3	58.4	69.1	72.0	65.8	7.1	31.7	65.8	35.2	61.3	43.2
Few-shot: exemplars with reasoning	GPT-4.1	47.3	55.6	39.0	52.2	59.4	55.6	67.5	56.1	61.3	56.6	70.9	58.2	75.1	70.8	67.3	6.1	31.6	65.8	36.5	54.6	39.6
	Gemini-2.0-Flash	51.0	58.1	**43.9**	54.4	62.3	58.1	65.8	55.0	59.9	57.2	74.2	59.3	68.5	68.9	64.0	**9.8**	**35.5**	70.6	**40.0**	60.3	**47.4**
	Claude-3.7-Sonnet	49.3	55.5	43.1	58.4	52.9	55.5	68.0	45.0	54.1	61.7	60.5	57.3	73.6	55.2	58.2	9.1	34.6	**70.8**	38.2	64.1	41.6
RAG: articles	GPT-4.1	46.4	53.3	39.5	47.0	61.6	53.3	65.2	62.4	63.8	53.1	71.1	55.9	70.2	70.2	66.4	6.8	32.4	64.4	36.5	54.0	43.0
	Gemini-2.0-Flash	45.7	50.0	41.5	56.4	44.9	50.0	**73.6**	42.9	54.2	61.0	57.2	53.2	76.8	55.0	59.4	6.9	32.8	67.2	35.6	66.4	40.1
	Claude-3.7-Sonnet	47.4	55.3	39.5	51.2	60.1	55.3	66.0	56.6	61.0	56.2	69.8	56.4	70.1	65.5	62.6	6.9	31.6	67.0	37.3	58.0	36.4
RAG: article summaries	GPT-4.1	46.7	52.8	40.5	49.1	57.2	52.8	62.7	53.4	57.7	54.9	66.6	55.4	68.8	63.0	61.1	7.2	33.4	67.4	36.8	55.4	42.9
	Gemini-2.0-Flash	45.7	49.0	42.3	52.0	46.4	49.0	60.2	39.2	47.4	57.1	55.1	50.0	66.7	49.9	51.1	6.4	33.0	66.8	38.7	**67.1**	41.7
	Claude-3.7-Sonnet	46.9	54.7	39.1	47.1	65.2	54.7	66.0	**66.7**	66.3	50.6	76.5	56.4	68.0	**74.7**	67.0	5.3	31.4	64.2	34.4	59.1	40.4
RAG: synthetic cases	GPT-4.1	47.1	56.3	37.9	50.0	64.5	56.3	62.4	58.7	60.5	55.6	74.2	59.8	68.3	68.3	65.1	4.8	29.2	64.0	34.6	54.7	40.3
	Gemini-2.0-Flash	48.9	58.2	39.7	55.2	61.6	58.2	66.9	54.5	60.1	59.7	68.6	58.0	72.9	65.1	62.8	4.8	30.9	66.7	35.7	59.4	40.5
	Claude-3.7-Sonnet	47.8	55.3	40.3	51.2	60.1	55.3	66.0	56.6	61.0	56.2	69.8	56.4	70.1	65.5	62.6	6.8	32.3	69.3	37.2	61.0	35.2
		Test dataset																			
Submission		Overall	Factual.	Relev.	SP_μ	SR_μ	SF_μ	LP_μ	LR_μ	LF_μ	SP_M	SR_M	SF_M	LP_M	LR_M	LF_M	BLEU	ROUGE-L	SARI	BERTScore	AlignScore	MEDCON
Exemplars as few-shot with Gemini-2.0-Flash		48.5	54.6	**42.5**	**62.6**	48.4	54.6	**65.9**	48.2	55.6	**67.6**	62.7	58.7	**71.4**	60.2	59.7	**6.3**	31.9	**67.7**	**37.0**	**68.7**	43.3
Exemplars as few-shot with GPT-4.1		48.6	**57.3**	39.8	61.4	**53.8**	**57.3**	64.7	**53.7**	**58.7**	65.7	**64.2**	**60.4**	70.2	**62.3**	**62.0**	4.2	29.6	64.6	33.8	63.7	43.1
Multistage few-shot prompting (Figure 1)		**49.1**	56.9	41.4	60.6	53.6	56.9	64.0	53.5	58.3	65.4	64.0	60.2	70.0	62.2	61.8	6.0	**32.1**	65.8	36.4	64.3	**43.6**

Table 1: Results on development and test sets. SP = Strict Precision, SR = Strict Recall, SF = Strict F1, LP = Lenient Precision, LR = Lenient Recall, LF = Lenient F1. Subscripts μ and M denote micro and macro respectively.

ble 1). Across all models, we observed consistently high macro recall but low micro recall. This suggests that while models can identify relevant evidence across different cases, they often fail to capture all necessary citations in longer notes with many sentences, indicating challenges in evidence completeness for long and complex cases.

7.1.2 Few shots outcome

GPT-4.1 with LLM-generated exemplars achieved the highest overall score of 51.5 on the development set, outperforming the reasoning-based few-shot approach. These exemplars notably improved factual recall, increasing the overall factuality score from 53.8 (baseline) to 61.1 without relying on external data. This highlights that well-structured, relevant examples can enhance the model's ability to cite appropriate evidence. In contrast, the reasoning-based few-shot setup achieved a lower overall score of 47.3, compared to 51.5 for few-shot prompting without reasoning. This suggests that explicitly including reasoning steps may not yield additional benefit in this task, and that the model may perform implicit reasoning more effectively when guided by concise, LLM-curated exemplars.

7.1.3 RAG: Full text vs. Article summary vs. Synthetic clinical cases

Among RAG variants with Gemini-2.0-Flash, top-5 synthetic cases yielded the best performance, achieving an overall score of 48.9 and improving factuality from 49.2 to 58.2 compared to the baseline. This suggests structured, case-like inputs better support clinical reasoning than unstructured text. RAG using full-text articles and summaries produced lower factuality scores (50.0 and 49.0, respectively). Although converting articles into cases improved alignment, these formats remained less effective than LLM-crafted exemplars, likely due to misalignment between retrieved content and the target question. Overall, the RAG approaches performed worse than the best few-shot LLM-generated exemplars. We hypothesized that the quality and relevance of in-context examples may be more important than retrieved knowledge.

7.2 Test dataset results

Based on the development set results, GPT-4.1 with LLM-generated exemplars as few-shot achieved the best overall performance. On the test set, GPT-4.1 demonstrated strong factuality (overall factuality = 57.3), while Gemini-2.5-Pro performed better in terms of relevance (overall relevance = 42.5). We leveraged both models by developing a multistage few-shot prompting pipeline without external data for our final submission, achieving an overall score of 49.1 (Table 1). This pipeline uses Gemini-2.5-Pro to generate 20 exemplar answers with citations from the development dataset. These 20 exemplars are used as in-context examples for GPT-4.1's initial answer generation on the test dataset. We then extract references from these initial answers. In the final stage, we input the test dataset and its corresponding retrieved references into Gemini-2.5-Pro to generate the final grounded answers (Figure 1).

8 Conclusion and Future Work

Our study demonstrates that few-shot learning with LLM-generated examples significantly improves EHR-grounded medical question answering. We achieved performance gains on the ArchEHR-QA 2025 benchmark without requiring model retraining or external knowledge sources. Models can leverage implicit patterns when guided by in-context learning demonstrations. Future work may explore example selection for ICL or demonstration strategy (Zhang et al., 2024; Huang et al., 2023), which can help improve the model's alignment with the ground truth. We can also improve the reference of clinical notes to achieve better recall.

Limitations

LLM-generated few-shot examples may incorporate subtle biases or inaccuracies that propagate through the system. Our implementation relies on underlying EHR data quality, which may vary in completeness and structure across clinical settings. In practice, real-world EMR heterogeneity amplifies these challenges: clinicians document information across free-text notes, scanned documents, and copied entries that vary widely in format, often include redundant or contradictory details, and fragment critical data. Moreover, we rely on proprietary model APIs with 19–20-shot prompts, which drive up computation time and latency and limit scalability in resource-constrained settings.

Despite strong benchmark performance, real-world deployment would require the validation of our prompting strategies on unstructured production EHR systems, incorporating robust NLP preprocessing (entity normalization, de-duplication) alongside human oversight to ensure clinical safety, data privacy, and appropriateness. We also need to prune exemplars, distill models, and conduct cost–benefit analyses to reduce inference time and API costs, all while upholding data privacy and regulatory compliance.

References

Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. 2024. Medhalu: Hallucinations in responses to healthcare queries by large language models. *Preprint*, arXiv:2409.19492.

Mohammad Alkhalaf, Ping Yu, Mengyang Yin, and Chao Deng. 2024. Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of Biomedical Informatics*, 156:104662.

Yaara Artsi, Vera Sorin, Eli Konen, Benjamin S Glicksberg, Girish Nadkarni, and Eyal Klang. 2024. Large language models for generating medical examinations: systematic review. *BMC Medical Education*, 24(1):354.

Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, Ulrich Keilholz, Ulf Leser, and Damian T. Rieke. 2023. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11):e2343689–e2343689.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Travis R. Goodwin, Dina Demner-Fushman, Kyle Lo, Lucy Lu Wang, Hoa T. Dang, and Ian M. Soboroff. 2022. Automatic question answering for multiple stakeholders, the epidemic question answering dataset. *Scientific Data*, 9:432.

John Holmes, James Beinlich, Mary Regina Boland, Kathryn Bowles, Yong Chen, Tessa Cook, George Demiris, Michael Draugelis, Laura Fluharty, Peter Gabriel, Robert Grundmeier, Clarence Hanson, Daniel Herman, Blanca Himes, Rebecca Hubbard, Charles Kahn, Jr, Dokyoon Kim, Ross Koppel, Qi Long, and Jason Moore. 2021. Why is the electronic health record so challenging for research and clinical care? *Methods of information in medicine*, 60.

Ziniu Huang, Jing Zhou, Guoxin Xiao, and Gong Cheng. 2023. Enhancing in-context learning with answer feedback for multi-span question answering. *arXiv preprint arXiv:2306.04508*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. 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. Published: 01 November 2023.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, and 6 others. 2025. Medical hallucination in foundation models and their impact on healthcare. *medRxiv*.

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 *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Mary M Lucas, Justin Yang, Jon K Pomeroy, and Christopher C Yang. 2024. Reasoning with large language models for medical question answering. *Journal of the American Medical Informatics Association*, 31(9):1964–1975.

Inc. Merck & Co. 2025. Msd manual consumer version. Retrieved May 3, 2025.

National Library of Medicine (US). 2025. Medlineplus. Bethesda (MD): National Library of Medicine; cited 2025 May 3.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7:322.

Yucheng Shi, Shaochen Xu, Tianze Yang, Zhengliang Liu, Tianming Liu, Quanzheng Li, Xiang Li, and Ninghao Liu. 2024. Mkrag: Medical knowledge retrieval augmented generation for medical question answering. In *Proceedings of the 2024 American Medical Informatics Association Annual Symposium (AMIA)*. Distinguished Paper Award.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, and 1 others. 2023. Toward expert-level medical question answering with large language models. *Nature Medicine*, 29:1–9.

Sarvesh Soni and Dina Demner-Fushman. 2025a. A dataset for addressing patient's information needs related to clinical course of hospitalization. *arXiv preprint*.

Sarvesh Soni and Dina Demner-Fushman. 2025b. Overview of the archehr-qa 2025 shared task on grounded question answering from electronic health records. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

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.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5:194.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuwei Zhang, Shi Feng, and Chenhao Tan. 2024. Demonstration selection for in-context learning

via reinforcement learning. *arXiv preprint arXiv:2412.03966*.

A Appendix

This appendix provides documentation of the prompts used in our multistage few-shot pipeline (Figure 1). The following sections describe the exact prompts, design rationale, and implementation considerations used throughout the system.

A.1 Prompt for Exemplar Generation

Figure B illustrates the prompt employed for generating exemplars. This stage utilizes the development dataset, which includes the patient narrative, patient question, clinician question, clinical note, and relevant sentence key. Gemini-2.5-Pro generates concise, citation-grounded answers in plain text, which are subsequently used as few-shot examples for downstream prompting.

A.2 Prompt for Initial Answers Generation

We use the prompt in Figure A to generate initial answers from the test dataset. Input components include the patient narrative, patient question, clinician question, and clinical note. The exemplars derived from the development data (as described in Section A.1) are incorporated into the prompt for initial responses with GPT-4.1.

LLM-Generated Exemplars as Few-shot

 # Examples

`{exemplars}`

 # To answer

Patient Narrative: `{patient_narrative}`
Patient Question: `{patient_question}`
Clinician Question: `{clinical_question}`
Clinical Note: `{clinical_note}`

Return your response in the format below strictly.

```
<answer>
Your answer based on the things you have seen in the Example Patient Narrative, Example Patient Question, Example Clinician Question, Example Clinical Note and the Example Answer. Please do not use a hyphen ('-') in the citation. List all the citations.
</answer>
```

Figure A: **Prompt used for LLM-generated exemplars as few-shot for final answers generation**

Exemplars and Final Answers Generation

You are a medical expert tasked with providing clear, accurate answers to medical questions based on relevant sentences from the clinical notes. Your response should be detailed, evidence-based, and reference specific points from the relevant sentences using the numbered citations. You are only allowed to use the relevant sentences to answer the question.

Example Patient Narrative:
I had severe abdomen pain and was hospitalised for 15 days in ICU, diagnosed with CBD sludge. Doctor advised for ERCP. My question is if the sludge was there does not any medication help in flushing it out? Whether ERCP was the only cure?

Example Patient Question:
My question is if the sludge was there does not any medication help in flushing it out? Whether ERCP was the only cure?

Example Clinician Question:
Why was ERCP recommended over a medication-based treatment for CBD sludge?

Example Clinical Note:
1: During the ERCP a pancreatic stent was . . .
2: However, due to the patient's elevated INR . . .
3: Frank pus was noted . . .
4: The Vancomycin was discontinued.
5: On hospital day 4 . . .
6: On ERCP the previous biliary stent . . .
7: As the patient's INR was normalized . . .
8: At the conclusion of the procedure . . .

Example Relevant Sentences: [1, 5, 6, 7]

Example Answer:
Medications can sometimes help in managing bile duct sludge, but in this case, ERCP was necessary... |1|... |5|... |6|... |7|.

Now, please provide a similar detailed answer for the following case:

Patient Narrative: `{patient_narrative}`
Patient Question: `{patient_question}`
Clinician Question: `{clinical_question}`
Clinical Note: `{clinical_note}`
Relevant Sentences: {relevant_sentences}

Answer Format:

```
<answer>
Use ALL of the relevant sentences to answer the question. Make sure to answer the question based on the relevant sentences. See the example answer for the format (use the |sentence number| to reference).
</answer>
```

Note: Think about the question and relevant sentences carefully. You may reshuffle the sentences, but should not include any other content.

Figure B: **Prompt used for exemplars and final answers generation**

A.3 Prompt for Final Answers Generation

For the final answer generation stage, we reutilize the prompt shown in Figure B. However, instead of using the development sentence key, we provide the model with retrieved sentences cited in the initial answers. This configuration enables Gemini-2.5-Pro to generate a grounded, citation-supported response based on the test data and previously extracted evidence.