LAILab at ArchEHR-QA 2025: Test-time scaling for evidence selection in grounded question answering from electronic health records

Tuan-Dung Le[1,2], **Thanh Duong**[1,2], **Shohreh Haddadan**[1],
Behzad Jazayeri[1], **Brandon Manley**[1], **Thanh Q. Thieu**[1,2]
[1]Moffitt Cancer Center and Research Institute, USA
[2]University of South Florida, USA

{tuandung.le , thanh.duong, shohreh.haddadan, behzad.jazayeri, brandon.manley, thanh.thieu}@moffitt.org

Abstract

This paper presents our approach to the ArchEHR shared task on generating answers to real-world patient questions grounded in evidence from electronic health records (EHRs). We investigate the zero-shot capabilities of general-purpose, domain-agnostic large language models (LLMs) in two key aspects: identifying essential supporting evidence and producing concise, coherent answers. To this aim, we propose a two-stage pipeline: (1) evidence identification via test-time scaling (TTS) and (2) generating the final answer conditioned on selected evidences from the previous stage. Our approach leverages high-temperature sampling to generate multiple outputs during the evidence selection phase. This TTS-based approach effectively explores more potential evidences which results in significant improvement of the factuality score of the answers.

1 Introduction

Large language models (LLMs) tuned with reinforcement learning from human feedback (RLHF), have transformed automatic question answering (QA) systems, leading to their widespread adoption in various domains. In clinical settings, QA systems have been used to answer health-related inquiries (Demner-Fushman et al., 2020) which require medical domain knowledge. Patient-specific QA, more critically, require grounding responses in evidence extracted from electronic health records (EHRs) to ensure factual accuracy and reliability. Training and fine-tuning of clinical-specific LLMs have been shown to outperform general models on NLP tasks, including patient-specific QA (Lehman et al., 2023). However, this approach faces several significant challenges. First, task-specific clinical data is often scarce and difficult to obtain due to strict privacy regulations and patient safety concerns. Second, manual expert annotation of such data is prohibitively expensive. Most critically, even when clinical datasets are de-identified, there remains a non-trivial risk of inadvertently disclosing protected health information (PHI) through model training and deployment (Das et al., 2025) specifically in real-world applications where models are accessible externally such as patient portals. These constraints, coupled with the increasing zero-shot capabilities of LLMs, motivate an alternative paradigm: leveraging general-purpose domain-agnostic LLMs and elicit their domain-specific knowledge and reasoning abilities at inference time. This approach known as test-time scaling (TTS) offers a promising path toward mitigating data scarcity, reducing annotation costs, improving robustness to input variability, and minimizing privacy risks in clinical NLP applications in real-world settings (Zhang et al., 2025).

In this paper, we present a TTS-based solution to the ArchEHR Shared Task (Soni and Demner-Fushman, 2025b). We argue that TTS is particularly well-suited for this task due to limited availability of annotated training data and the method's practicality in real-world deployment scenarios, such as integration into patient portals. We propose a two-stage pipeline methodology consisting of evidence identification followed by answer generation. In the first stage, we employ a parallel TTS strategy by generating multiple outputs at a high temperature and selecting frequently predicted sentences as essential evidence. In the second stage, we prompt the model to generate concise and grounded answers conditioned on the selected evidence, using different prompting strategies to optimize response quality.

2 Task Description

The **ArchEHR-QA 2025** shared task aims at automatically providing answers to real-world patient questions grounded in evidence from EHRs (Soni and Demner-Fushman, 2025b). The dataset con-

BioNLP 2025 Shared Tasks, pages 75–80
August 1, 2025 ©2025 Association for Computational Linguistics

sists of 20 cases in the development set and 100 in the test set (Soni and Demner-Fushman, 2025a). Each case includes patient question, clinician-rewritten version, and excerpts from patients' clinical notes. Each sentence from the excerpt is manually labeled as *essential*, *supplementary*, or *not relevant*, indicating the relevance of the sentence to the answer. Systems are evaluated on two criteria: factuality and relevance. Overall factuality is assessed using strict micro F1, where only essential evidence sentences are considered relevant, with manual annotations as reference labels. Automated relevance is measured by comparing generated answers to reference texts, which include patient narrative, clinician question, and ground-truth evidence sentences. Relevance metrics are BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), SARI (Xu et al., 2016), BERTScore (Zhang et al., 2020), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023). The final leaderboard score averages the overall factuality score and the normalized average of all automated relevance metrics. The organizers also conduct additional post-challenge evaluations, including relevance comparisons to clinician-written answers and manual assessments, offering a more comprehensive view of system performance (Soni and Demner-Fushman, 2025b).

3 Approach

3.1 Overview

To address the challenges posed in low-resource settings given only 20 cases in development set, we leverage the strong zero-shot capabilities of LLMs. Our preliminary experiments in prompting LLMs to directly generate answers using corresponding citations result in high variability across runs and inconsistent sets of cited evidence generated at each run by the same prompt. This method also often leads to low overall factuality scores. These initial findings align with the baseline scores reported by the organizers using a similar strategy.

To address this limitation towards a more reliable patient-specific QA system grounded in evidences from note excerpts, we propose a two-stage prompting strategy. In the first stage, we apply parallel test-time scaling to identify a broader set of potentially essential evidence sentences. In the second stage, we generate the final answer conditioned on the evidence selected during the first stage.

3.2 Stage 1: Evidence identification

The goal of this stage is to identify essential sentences from the note excerpt to serve as evidence to answer the patient's question. Given a clinical note consisting of sentences s_i for $i = 1, 2, \ldots, N_{sent}$ where N_{sent} is the total number of sentences in the note, we prompt a LLM to generate a list of relevant sentence indices i. We apply a zero-shot chain-of-thought prompting strategy (Wei et al., 2022), using the following prompt:

```
Given a clinical note and a patient's question,
identify the sentence indices that provide evidence to
answer the question. Each sentence in the clinical
note is indexed. Return only the relevant sentence
indices as a comma-separated list.

Clinical note: ...
Patient question: ...

Think step by step before finalizing your answer.
Provide your final answer within \boxed{{}}.
```

We extract a list of sentence indices from each model-generated output, representing the sentences identified as essential. To encourage diversity in evidence selection, we sample multiple candidate outputs by varying the decoding temperature. A lower temperature (e.g., 0) results in more deterministic outputs, while a higher temperature (e.g., 0.6 or 1.0) increases randomness, allowing the model to explore more candidate solutions (Renze, 2024). We prompt the model once using temperature 0 (greedy decoding), 64 times with a temperature of 0.6, and either 128 or 256 times with temperature a of 1 to encourage diverse output generation. Let c_i denote the number of times sentence s_i is predicted as essential across all runs. A sentence is included in the final evidence set if $c_i \geq t$, where t is a threshold in the range $[1, N_{gen}]$ and N_{gen} is the total number of generations.

For this stage, we employ two open instruction-tuned LLMs: Qwen2.5-32B-Instruct (Yang et al., 2024) and LLaMA-3.3-70B-Instruct (Grattafiori et al., 2024). We investigate the effectiveness of three question variations provided in the dataset: patient narratives, patient questions, and clinician questions. Results on the development set indicate that prompts with solely patient narratives as input consistently achieve the highest performance. Accordingly, all prompts in our experiments use only patient narrative as input.

3.3 Stage 2: Answer generation

We prompt the LLM to generate the final answer using the essential sentence indices identified in

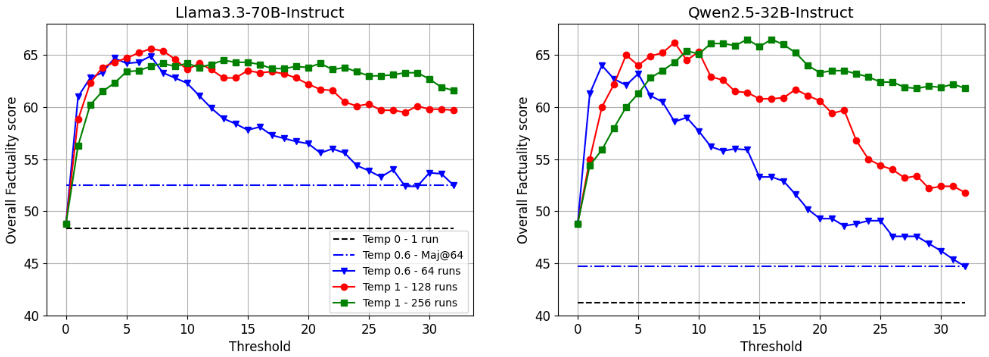

Figure 1: Test-time scaling improves factuality score on development set.

the previous stage with the following prompt:

```
Given a clinical note, a patient’s question and a list
of sentence indices that represent the essential
supporting evidence, write a 5-sentence (fewer than
100 words) answer that addresses the patient’s
concern. Each sentence must end with the evidence
indices immediately after the period, in this format:
"The treatment was successful.|1,2|\n"

You must cite all essential indices in the answer. Do
not introduce any information that is not grounded in
the clinical note. To ensure high-quality answer,
reuse as much phrasing and sentence structure from the
clinical note as possible.

Clinical note: ...
Patient question: ...
Essential sentences: <list of sentence indices from
stage 1>
```

We conduct an ablation study by varying the instruction components to evaluate their impact on the overall score. Specifically, we experiment with constraints such as allowing free-form generation, limiting the answer length to a fixed number of sentences or words, and encouraging the model to reuse phrasing, sentence structure, or exact evidence sentences from the clinical note.

In this stage, we experiment with Gemini-2.0-flash(Google, 2024) and Gemini-2.5-pro-preview(Google, 2025)[1], as these models more reliably follow instructions and consistently generate answers in the required submission format, whereas the open-source LLMs used in Stage 1 occasionally fail to meet these criteria.

4 Results and Discussions

4.1 Dev performance

Figure 1 shows the performance of the evidence identification stage. These results indicate that generating multiple outputs with higher temperatures and setting lower selection thresholds consistently improves factuality scores. With a temperature of 1.0, Llama-3.3-70B-Instruct achieves a factuality score of 65.4 using a threshold of 7 over 128 runs, while Qwen2.5-32B-Instruct achieves the highest score of 66.5 with thresholds of 14 and 16 over 256 runs. This approach outperforms both single-pass greedy decoding and self-consistency with majority voting (Wang et al., 2022).

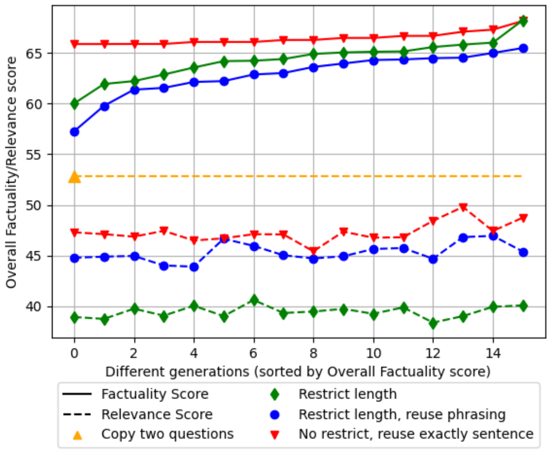

Figure 2: Factuality and relevance scores of answer generation strategies on the development set, evaluated over 16 runs using Gemini-2.0-Flash at temperature 0.6.

For each case, we prompt the model 16 times with different configurations using the best evidence set identified in Stage 1. Figure 2 presents the performance of our answer generation strategies across the 16 runs. When the model is restricted to generate answers with a maximum of 5 sentences and fewer than 100 words, the model achieves an overall relevance score of approximately 40, with an average output length of approximately 80 words on the development set.

It’s important to note that despite explicitly in-

[1] These models are accessed via Vertex AI, the platform recommended by PhysioNet for responsible MIMIC data use.

#ID	Stage 1	Stage 2	Leaderboard			Post-challenge re-evaluation			
	t	Answer generation settings	Ovr.	Fact.	Auto Rel.	Human Ovr.	Fact.	Auto Rel.	Human Rel.
1	16	Gemini-2.5-pro-preview 5 sentences, $\leq$ 100 words reuse phrasing and sentence structure	48.2	59.2	37.3	**43.1**	**53.8**	38.0	**32.4**
2	16	Gemini-2.0-flash 5 sentences, $\leq$ 100 words add unused citations to last sentence reuse exact sentence when possible	49.7	59.1	40.3	42.6	53.5	41.5	31.7
3	14	Gemini-2.0-flash no limit reuse exact sentence when possible	**51.0**	**60.4**	**41.6**	41.5	53.3	**42.0**	29.6

Table 1: Details of our three submissions on the test set. Leaderboard scores are based on initial relevance labels and concatenated evidence sentences from the clinical notes, while post-challenge re-evaluation scores use reconciled relevance labels and clinician-written reference answers.

structing the model to include all essential sentences from stage 1, LLMs often omit or introduce citations outside the provided list, leading to variability across runs. The factuality score varies by up to 6 points, while the relevance score remains relatively stable. Removing length constraints improves citation consistency, with the model more reliably preserving the majority of the evidence sentences identified in the previous stage.

We observe that automated relevance metrics favor answers that closely align with the reference, which integrates information from the patient narrative, clinician questions, and ground-truth essential sentences. Prompting the model to reuse phrasing or directly incorporating sentences from the clinical note consistently boosts relevance scores to the 45–47 range. Further improvements are achieved by directly copying sentences from the identified evidence and ordering them based on importance or model confidence to prioritize key information within the first 75 words of the generated response. Moreover, using the patient narrative and clinician question directly as the answer (or appending them to the beginning of the answer) yields a relevance score of 52.9, significantly improving all automated relevance scores, except for SARI score due to copying questions. However, we refrain from adopting these direct copy strategies in our final submission, as they diverge from the objective of the challenge, which emphasize generating coherent responses.

A medical expert at our institute provides answers for the development set based on the annotated essential sentences. Their responses yield an average relevance score of 27.2 with an average length of 54 words, excluding case 16, where our expert notes that the clinical note lacks relevant evidence to answer the patient's question.

4.2 Test submissions

Details of our three test submissions are shown in Table 1. We run Qwen2.5-32B-Instruct 256 times and select essential sentences using thresholds of 14 or 16, chosen based on development set performance. For the first submission, we use Gemini-2.5-pro-preview, which includes all essential sentences within 5 sentences likely due to its stronger reasoning capabilities. The other two use Gemini-2.0-flash to boost automated relevance scores.

Post-challenge re-evaluation based on reconciled relevance labels results in factuality scores dropping by up to 7.1 points, while automated relevance scores varies only slightly, increasing by at most 1.2 points. This aligns with our development set observations and highlights the limitations of automated relevance metrics. Mitigating the limitations of automated relevance scores, the organizers evaluated human relevance by comparing our answers with clinician-written reference answers. Interestingly, human relevance scores often diverged from automated ones, favoring shorter responses with less verbatim replication of the evidence sentences.

5 Related Work

Extractive question answering—a task closely related to grounded question answering—aims to extract patient-specific answer spans from clinical notes in response to clinical queries. Recent approaches have leveraged large language models (LLMs) to address this challenge through a variety of techniques. Fine-tuning language models such as ClinicalBert for sequence generation (Moon et al., 2023) and sequence labeling(Yue et al., 2021) tasks was used for extractive QA from unstructured EHR notes. Hamidi and Roberts (2023) experiment prompting ChatGPT 3.5 and Claude and report a manual evaluation of accuracy, relevance, comprehensiveness, and coherence on a set of patient-specific questions. Lehman et al. (2023) evaluate the performance of various clinical domain specific LLMs with different sizes ranging from 220M to 175B parameters, and use in context learning (ICL) for extractive QA on a dataset on radiology reports (Soni et al., 2022). Their results demonstrate that fine-tuning clinical domain specific models outperform ICL methods on extractive QA.

6 Conclusion

Zero-shot prompting of large language models for patient-specific question answering—grounded in clinical notes—results in inconsistent evidence selection, leading to lower factuality scores. Parallel scaling strategy at test-time mitigates this problem in a low-resource setting. We experiment with generating multiple outputs at higher temperatures and selecting frequently predicted sentences as essential evidence which improves factuality score of evidence identification. We then generate answers conditioned on the selected evidence, and further enhance relevance by engineering the prompt to align the answer to the question while preserving coherence.

Limitations

Our proposed approach has several limitations. First, applying TTS by generating multiple outputs increases computational cost and latency. We run the Qwen2.5-32B-Instruct model 256 times on 4 H100 GPUs to identify evidence, averaging 4 seconds per case, followed by answer generation with Gemini-2.0-Flash via API, which takes an additional 1 second. Due to the cost, we avoid using API-based models for evidence selection and instead rely solely on open-source instruction-tuned LLMs. Exploring more efficient TTS methods with recent open-weight reasoning models such as DeepSeek-R1(Guo et al., 2025) and Qwen3(Yang et al., 2025) is a promising direction for future work. Second, the frequency-based evidence selection is tuned on a small development set of 20 examples, which may not generalize well to unseen cases. Third, while the use of API-based models for answer generation is acceptable for this shared task, it may not be feasible or allowed in real-world clinical settings due to privacy and regulatory constraints. Finally, the answer quality is sensitive to prompt design in the second stage, with minor phrasing changes often leading to significant output variability.

References

Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2025. Security and privacy challenges of large language models: A survey. *ACM Comput. Surv.*, 57(6).

Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2020. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association*, 27(2):194–201.

Google. 2024. Gemini-2.0-flash-001. https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-2.0-flash-001.

Google. 2025. Gemini-2.5-pro-preview-03-25. https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-2.5-pro-preview-03-25.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Alaleh Hamidi and Kirk Roberts. 2023. Evaluation of ai chatbots for patient-specific ehr questions. *arXiv preprint arXiv:2306.02549*.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? In *Conference on health, inference, and learning*, pages 578–597. PMLR.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

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.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA. Association for Computational Linguistics.

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.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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.

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

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.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. 2025. What, how, where, and how well? a survey on test-time scaling in large language models. *arXiv preprint arXiv:2503.24235*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.