Neural at ArchEHR-QA 2025: Agentic Prompt Optimization for Evidence-Grounded Clinical Question Answering

Sai Prasanna Teja Reddy[1], Abrar Majeedi, Viswanatha Reddy Gajjala, Zhuoyan Xu, Siddhant Rai, Vaishnav Potlapalli

[1]University of Chicago
bogireddyteja@uchicago.edu

Abstract

Automated question answering (QA) over electronic health records (EHRs) can bridge critical information gaps for clinicians and patients, yet it demands both precise evidence retrieval and faithful answer generation under limited supervision. In this work, we present *Neural*, the runner-up in the BioNLP 2025 ArchEHR-QA shared task on evidence-grounded clinical QA. Our proposed method decouples the task into (1) sentence-level evidence identification and (2) answer synthesis with explicit citations. For each stage, we automatically explore the prompt space with DSPy's MIPROv2 optimizer, jointly tuning instructions and few-shot demonstrations on the development set. A self-consistency voting scheme further improves evidence recall without sacrificing precision. On the hidden test set, our method attains an overall score of 51.5, placing second stage while outperforming standard zero-shot and few-shot prompting by over 20 and 10 points, respectively. These results indicate that data-driven prompt optimization is a cost-effective alternative to model fine-tuning for high-stakes clinical QA, advancing the reliability of AI assistants in healthcare.

1 Introduction

Automatically generating answers to patients' medical questions using information from their electronic health records (EHRs) poses significant challenges, but also offers substantial potential for improving clinical communication and patient engagement (Soni and Demner-Fushman, 2025b). The ArchEHR-QA 2025 shared task directly targets this problem by providing patient questions alongside excerpts from clinicians' notes, and requiring systems to generate grounded responses that explicitly cite the supporting sentences.

Recent advances in Large Language Models (LLMs) have shown promising results in the domain of answering clinical questions based on unstructured patient notes (Singhal et al., 2025). However, fine-tuning LLMs for answering clinical questions based on unstructured patient notes is constrained by the limited availability of supervised clinical data, which increases the risk of overfitting. Consequently, prompt-based methods offer a practical and cost effective solution. Despite the advantages, prompt engineering comes with its own set of challenges (Karayanni et al., 2024). Crafting effective prompts for complex tasks often requires expert effort and iterative refinement. This difficulty is amplified in the clinical domain, where identifying the correct evidence from lengthy medical notes is critical for accurate answers. Prior studies have explored manual prompt designs and chain-of-thought cues for medical QA (Tai and Tannier, 2025), yet these ad-hoc methods may not yield optimal performance. Automated prompt optimization techniques (Wang et al., 2023) offer a systematic alternative, but often treat each task holistically and may not incorporate domain expertise effectively.

In this work, we introduce a two-stage LLM pipeline for clinical question answering that explicitly separates evidence identification and answer generation. In each stage, prompts are automatically optimized using the MIPROv2 optimizer from DSPy (Khattab et al., 2021, 2024). The first stage is dedicated to identifying the relevant information within the clinical note, while the second stage leverages this information to generate a precise and contextually appropriate answer. By decomposing the task into these two distinct stages, it becomes possible to define clear, stage-specific evaluation objectives, namely, F1 score for evidence retrieval performance and the mean of word limit score, citation format score, BLEU, ROUGE, SARI, BERT, ALIGN, and MEDCON scores for answer quality metrics. This decomposition also enables the use of optimization algorithms to systematically search for prompts that maximize these objectives. To further improve reliability, we integrate a self-

BioNLP 2025 Shared Tasks, pages 104–109
August 1, 2025 ©2025 Association for Computational Linguistics

consistency (Wang et al., 2022) approach in the evidence retrieval stage: the model is run multiple times, and a majority vote over the outputs determines the final cited sentences, reducing variability and errors.

In summary, our contributions are:

- **Decomposed Prompt Optimization Framework:** We propose a two-stage pipeline that modularizes clinical QA, enabling distinct and targeted prompt optimization for evidence retrieval and answer generation, a paradigm shift from monolithic optimization approaches.
- **Systematic Instruction Space Exploration:** We demonstrate the efficacy of leveraging advanced optimizers like MIPROv2 to discover high-performing, task-specific prompt configurations from limited development data, enhancing both performance and reproducibility.
- We perform a rigorous evaluation on an expert-annotated clinical QA dataset, demonstrating that our prompt-optimized pipeline yields significant improvements in factual accuracy and answer relevance compared to established baselines, underscoring its effectiveness for reliable clinical-QA.

2 Related Work

Clinical QA: Developing QA systems for clinical data has long been an interest in biomedical NLP. Earlier datasets like emrQA (Pampari et al., 2018) generated large-scale QA pairs from electronic medical records by repurposing annotations, but these often contained synthetic questions or required mapping to structured outputs. Recent research has shown that large LLMs can achieve near-expert performance on medical QA benchmarks (Singhal et al., 2025).

Prompt Optimization: There is a growing interest in automated prompt search or optimization. More recently, methods such as APE (Zhou et al., 2022) and OPRO (Yang et al., 2023) treat prompt design as a black-box optimization problem, iteratively refining prompts by evaluating model outputs. MIPRO (Opsahl-Ong et al., 2024) extends this idea to multi-stage LLM programs, jointly optimizing the instructions and demonstration examples of each module in a pipeline. Our work leverages the latest optimizer, MIPROv2 (Opsahl-Ong et al., 2024), which uses a combination of prompt proposal and Bayesian search to find high-performing prompts efficiently.

Self-Consistency: Large LLMs can produce variable outputs given the same prompt, especially under chain-of-thought reasoning. The self-consistency decoding strategy (Wang et al., 2022) addresses this by sampling multiple outputs and choosing the result that is most consistent across samples.

3 Methodology

Our method draws on a human-inspired decoupling strategy, separating evidence gathering from solution formulation. In Stage 1, we identify relevant resources analogous to conducting a web search or literature review by retrieving key sentences. In Stage 2, we frame the final solution by synthesizing insights from the retrieved facts. We operationalize this intuition as a modular, two-stage pipeline tailored to clinical QA.

Consider each clinical note excerpt is segmented into individual sentences $s_1, s_2, \ldots, s_n$, and each sentence s_i is annotated with a label $y_i \in$ (essential, not-relevant, supplementary. The label indicates whether s_i contains information essential for answering a given patient/clinician question q. This sentence-level annotation forms the basis of Stage-1, while Stage-2 uses the content of the essential sentences (post consistency testing) to produce the final answer a_{gen}.

3.1 Sentence-Level Essentiality Classification

For a question–note pair let

$$Y^+ = \{ i \mid y_i = 1 \}, \text{and } \hat{Y}^+ = \{ i \mid \hat{y}_i = 1 \},$$

denote, respectively, the indices of *gold-standard essential* sentences and the indices predicted essential by the model. We begin with a manually crafted prompt that presents the question q and the sentence sequence $\{s_1, \ldots, s_n\}$ and requests a binary relevance label for every sentence in addressing the q.

Prompt-Optimization Objective (Stage-1): We invoke the **MIPROv2** to optimize the prompt. Treating the instruction text (and any embedded demonstrations) as discrete parameters $P \in \mathcal{P}$, MIPROv2 iteratively: (i) proposes a candidate

prompt P, (ii) applies the fixed LLM to the training set, and (iii) updates P so as to *maximize* the sentence-level $F_1(Y^+, \hat{Y}^+)$. By searching this space of instructions and few-shot exemplars, the optimizer converges on a prompt P^* that elicits labels with markedly higher precision and recall, thereby yielding a more reliable evidence set for Stage 2.

Self-Consistency Voting: To improve the reliability of Stage 1, we apply a *self-consistency voting* scheme: the classifier is executed $R = 5$ times on the same $(q, \{s_i\})$ input, each run differing only in its stochastic seed. Let $\hat{y}_i^{(r)} \in \{0, 1\}$ be the binary prediction for sentence s_i in run r (1 = essential). The final label is obtained by majority vote,

$$v_i = \sum_{r=1}^{R} \hat{y}_i^{(r)}, \quad \hat{y}_i = \begin{cases} 1 & \text{if } v_i \geq \tau = \lceil R/2 \rceil, \\ 0 & \text{otherwise,} \end{cases}$$

This aggregation suppresses spurious single-run errors and retains sentences identified as essential by at least three of the five passes, thereby reducing variance and boosting the expected F_1 of the evidence selection step.

3.2 Answer Generation from Essential Sentences

Let q be the input question and let $E = \{ s_i \mid \hat{y}_i = 1 \}$ denote the set of sentences that Stage 1 predicted as *essential*. Given the pair (q, E), Stage 2 must produce a concise natural-language answer a_{gen} that (i) directly addresses q, (ii) contains at most 75 words, and (iii) cites the supporting sentences in E using the required parenthetical notation. We initialise Stage 2 with a hand-written prompt template and then invoke **MIPROv2** to optimize this template. Let P denote a prompt parameterised by its instruction wording and any embedded demonstrations, and let $g_\theta(\cdot\,; P)$ be the fixed LLM generator. Given an input pair (q, E) the model outputs $a_{\text{gen}} = g_\theta((q, E); P)$.

Prompt-Optimization Objective (Stage-2): The goal is to maximise the composite reward

$$\mathcal{R}(a_{gen}, a^*, E) = \underbrace{\mathbf{1}[|a_{gen}| \leq 75]}_{\text{length}} + \underbrace{\mathbf{1}[\text{format}(a_{gen}, E)]}_{\text{citations}} + \underbrace{\frac{1}{6} \sum_{m \in \mathcal{M}} m(a_{\text{gen}}, a^*)}_{\text{surface \& semantic quality}}$$

where a^* is the reference answer, $|\cdot|$ counts words, and

$$\mathcal{M} = \{\text{BLEU, ROUGE, SARI, BERT, Align, MEDCON}\}.$$

The indicator terms enforce hard constraints on length and citation format, while the mean of the six metrics rewards lexical overlap, semantic fidelity, factual consistency, and medical–concept coverage.

Search Procedure: MIPROv2 explores the discrete prompt space $\mathcal{P}$ by iteratively proposing candidate prompts, evaluating them on a validation set, and selecting

$$P^* = \underset{P \in \mathcal{P}}{\text{argmax}} \; \mathbb{E}_{(q,E,a^*)} \Big[\mathcal{R}\big(g_\theta((q, E); P), a^*, E\big) \Big]$$

The optimal prompt P^* consistently elicits answers that are succinct, properly cited, and of high quality according to all surface–level and semantic metrics, thus satisfying the Stage 2 requirements.

4 Experimental Setup

Dataset: We evaluated our system on the ArchEHR-QA 2025 dataset (Soni and Demner-Fushman, 2025a). This dataset contains 120 question-note cases derived from MIMIC-III/IV clinical notes. Each case includes a patient question (often a layperson's phrasing) and a clinician-rewritten question focusing on the key medical query, along with a relevant excerpt from the patient's EHR notes. The notes are annotated with sentence numbers and labels indicating relevance ("essential," "supplementary," "not relevant") to the question. The official split provides 20 cases as a development set and 100 cases as a test set . We used the 20 development cases for prompt optimization and for all ablations. Final results on the test set were obtained via the Codabench submission system.

Evaluation Metrics: Following the official ArchEHRQA shared task protocol, we evaluate each submission along two complementary axes: *Factuality* and *Relevance*, which help capture evidence faithfulness and response quality. *Factuality* is quantified by matching the set of note sentences cited by the model against expert-annotated evidence and computing precision, recall, and F1. We report a **strict** variant that counts only *essential*

Table 1: Evaluation of participants on factuality and relevance metrics. **Bold** indicates the best performance in each column, underlined the second best. Here P^S, R^S, F_1^S denote micro-averaged *strict* precision, recall and F1; P^L, R^L, F_1^L denote micro-averaged *lenient* precision, recall and F1; AVG_{fact} and $\text{AVG}_{\text{relev}}$ are the official "Overall Factuality" and "Overall Relevance" scores, and "Overall" is the combined score. Abbreviations: R.L.= ROUGE-Lsum, B.S. = BERTScore, A.S. = AlignScore, M.C. = MEDCON .

Model	Factuality							Relevance							Overall
	P^S	R^S	F_1^S	P^L	R^L	F_1^L	AVG_{fact}	BLEU	R.L.	SARI	B.S.	A.S.	M.C.	$\text{AVG}_{\text{relev}}$	
DMISLab 🏆	57.9	59.3	58.6	61.2	59.2	60.2	58.6	**14.3**	**46.5**	36.7	**53.9**	**92.4**	**49.3**	**48.8**	**53.7**
Ours 🏆	55.4	63.8	59.3	58.4	63.7	60.9	59.3	8.5	34.1	**73.1**	39.1	67.3	40.0	43.7	51.5
LAILab 🏆	56.0	**65.5**	**60.4**	59.7	**66.0**	**62.7**	**60.4**	6.5	32.7	69.2	37.4	65.3	38.4	41.6	51.0
LAMAR	60.6	53.6	56.9	64.0	53.5	58.3	56.9	6.0	32.1	65.8	36.4	64.3	43.6	41.4	49.1
ssagarwal	68.8	36.2	47.5	71.7	35.6	47.6	47.5	4.7	31.1	70.0	36.9	74.9	38.0	42.6	45.0
Few-Shot	71.2	38.2	49.8	74.5	37.8	50.2	49.8	1.7	25.5	53.9	28.7	54.5	39.7	34.0	41.9
Zero-Shot	**71.6**	21.9	33.6	**77.0**	22.3	34.6	33.6	0.1	15.2	47.8	20.5	57.7	25.6	27.8	30.7

citations and a **lenient** variant that also accepts *supplementary* evidence, following the task guidelines. *Relevance* is evaluated as the arithmetic mean of complementary surface and semantic level metrics: BLEU (Papineni et al., 2002), ROGUE (Lin, 2004), SARI (Xu et al., 2016), BERTScore (Zhang et al., 2020), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023).

Baselines: To gauge the performance of our prompt optimization approach, we compare it against two baselines:

- **Zero-Shot Prompting:** A single, succinct instruction per stage. This reflects the common practice of "plug-and-play" prompting without any exemplars.
- **Few-Shot Prompting:** Adds two manually selected demonstrations to each stage's prompt but preserves the terse directive style. This isolates the value of exemplars alone, without optimization.

5 Results

Table 1 presents the comparative performance of our system alongside competing submissions on the ArchEHR-QA 2025 test set. Our approach ranked **second overall**, achieving a combined score of **51.5**, with individual scores of **59.3** for factuality and **43.7** for relevance. Crucially, our system maintained **consistently high performance across all evaluation axes**, in contrast to other systems that exhibited strong performance on isolated metrics but lacked robustness overall. We observe a **substantial margin of improvement over baseline prompting strategies**: our method outperforms the **zero-shot** and **few-shot** variants by approximately **20 and 10 points**, respectively, on the overall score. These gains underscore the effectiveness of *automated prompt optimization*, which systematically discovers high-performing instructions and demonstrations tailored to each stage of the QA pipeline. Moreover, our system's relative stability across metrics—including both surface-level (BLEU, ROUGE, SARI) and semantic (BERTScore, AlignScore, MEDCON) relevance measures—suggests that **prompt optimization not only improves individual metrics but also contributes to the holistic quality and trustworthiness of generated answers**. These findings affirm our central claim: that prompt optimization is not merely a heuristic tuning step, but a principled and impactful method for enhancing LLM-based clinical QA systems.

6 Conclusion

We propose a two-stage approach for clinical question answering on medical notes, leveraging DSPy's MIPROv2 optimizer to autonomously fine-tune prompts for each stage. In Stage 1, the method extracts essential evidence from the notes by optimizing the prompt to maximize the evidence F1 score. In Stage 2, the system generates answers by optimizing a prompt based on a composite metric incorporating several metric (word limit score, citation format score, BLEU, ROUGE, etc.), yielding concise, structured, and clinically reliable response. This prompt-optimized pipeline demonstrates substantial improvements over baselines, highlighting the efficacy of prompt optimization within a modular LLM framework. The results suggest that prompt engineering can transit from heuristic practice to data-driven optimization process, identifying high-performing prompts tailored to specific tasks. For medical question answering systems,

this advancement enhances both evidence retrieval and answer trustworthiness, representing a significant step toward the development of reliable AI assistants for clinicians and patients.

Future research directions include integrating web search agents to retrieve external medical knowledge absent from clinical notes, further enriching the capabilities and completeness of automated clinical QA systems.

7 Limitations

Despite strong performance on the ArchEHR-QA benchmark, our two-stage prompt-optimized framework faces limitations rooted in both data and model design. The curated and annotated EHR excerpts used for evaluation do not reflect the messiness of real-world clinical notes, which often suffer from incompleteness, inconsistency, and institutional variability; this makes generalization across healthcare settings difficult, especially given the lack of standardization and privacy restrictions on accessing realistic data. Furthermore, the model has not been domain-adapted and relies on a generic tokenizer, potentially missing specialized medical vocabulary crucial for understanding nuanced queries. The modular two-step process, while flexible, introduces latency and risk of compounding errors, especially as the size of the candidate space in MIPROv2 grows. This reranker also depends heavily on metrics like BLEU, which can reward surface-level similarity over true semantic alignment and are sensitive to the distribution of training data. Together, these factors raise concerns about both scalability and the quality of alignment, even when evaluation scores appear strong.

8 LLM Settings

In both stages of our pipeline—sentence-level evidence identification and answer synthesis—we employ the GPT-4.1 model accessed via the OpenAI API. To accommodate the extensive clinical context and few-shot demonstrations during prompt optimization, we allocate a maximum context window of 10,000 tokens. All prompt-optimization experiments (i.e., MIPROv2's evaluation of candidate prompt templates and few-shot exemplars) are conducted with a low-variance decoding strategy, setting the temperature to 0.3. This relatively "cold" sampling regime promotes determinism, ensuring that our optimizer receives consistent feedback on prompt efficacy as measured by evidence-retrieval F1 or composite relevance metrics.

For the self-consistency mechanism in Stage 1, we leverage stochastic sampling to capture the model's latent uncertainty. Specifically, we issue R = 5 independent generations per question–note pair, each sampled at temperature 0.7. A majority-vote over these five runs determines the final label for each sentence, suppressing spurious outliers while preserving genuinely informative evidence. All other decoding parameters (e.g., top-p, frequency and presence penalties) are held at their API defaults, isolating temperature and context length as the principal levers in our experimental configuration.

9 Prompts and Code Availability

To promote transparency and reproducibility, we release all manual and optimized prompt templates, together with our full pipeline implementation at our GitHub repository[1].

References

Nader Karayanni, Aya Awwad, Chein-Lien Hsiao, and Surish P Shanmugam. 2024. Keeping experts in the loop: Expert-guided optimization for clinical data classification using large language models. *arXiv preprint arXiv:2412.02173*.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Relevance-guided supervision for OpenQA with ColBERT. *Transactions of the Association for Computational Linguistics*, 9:929–944.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.

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.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. *arXiv preprint arXiv:2406.11695*.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.

[1] https://github.com/ViswanathaReddyGajjala/ArchEHR-QA-Neural

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

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.

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.

Chi-en Amy Tai and Xavier Tannier. 2025. Clinical trial cohort selection using large language models on n2c2 challenges. *arXiv preprint arXiv:2501.11114*.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P Xing, and Zhiting Hu. 2023. Promptagent: Strategic planning with language models enables expert-level prompt optimization. *arXiv preprint arXiv:2310.16427*.

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*.

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.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

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. *arXiv preprint arXiv:2305.16739*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.