OptiMed at ArchEHR-QA 2026: GEPA Prompt Optimization and Multi-Agent Majority Voting for EHR-Grounded Question Answering

Feras AlMannaa[1], **Talia Tseriotou**[2], **Maria Liakata**[2,3]
[1]Independent Researcher [2]Queen Mary University of London [3]The Alan Turing Institute
{t.tseriotou, m.liakata}@qmul.ac.uk

Abstract

Despite the demonstrated promise of Large Language Models in medical question answering, existing work largely addresses closed-form, exam-style tasks and overlooks complex open-ended questions requiring reasoning over noisy, long clinical documents. In this work, we present our system, **OptiMed**, submitted to the ArchEHR-QA 2026 shared task on grounded clinical question answering over EHR notes. We combine GEPA, an evolutionary prompt optimization framework, with multi-agent majority voting across five diverse LLMs and a structured clinical abstraction strategy for question interpretation. OptiMed ranked 1st overall among teams completing all four subtasks with an average score of 52.0, achieving top AlignScore in both Question Interpretation and Answer Generation, reflecting strong factual grounding. GEPA optimization proved effective for structured tasks with sufficient development data, but failed to generalize on complex generative tasks under very limited number of supervisions. Multi-agent majority voting consistently lifted performance in evidence-oriented subtasks. Prompt analysis attributes GEPA's gains to role prompting and procedural decomposition and failures to over-specification under limited supervision.

Keywords: Clinical Question Answering, Electronic Health Records, Prompt Optimization, Multi-Agent Framework

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in Medical Question Answering (QA), achieving expert-level performance across academic medical knowledge benchmarks and real-world clinical judgment tasks (Nori et al., 2023; Liévin et al., 2024; Singhal et al., 2025). Despite their success, such evaluations remain confined to exam-style, closed-form QA, failing to capture the complexity of open-ended patient questions that require reasoning over noisy, lengthy clinical documents (Bardhan et al., 2024) and grounding in the patient's unique medical history.

Yet few works have addressed the effectiveness of LLMs on Clinical QA over patient-specific records (Ahsan et al., 2024; AlMannaa et al., 2025), with most efforts focusing on condition identification or implicit diagnosis over EHRs (Pan et al., 2025; Albassam et al., 2025). Such tasks are well-structured and limited in scope. Other work focuses on EHR QA dataset development, which is instrumental in enabling downstream applications (Bardhan et al., 2022; Kweon et al., 2024).

With the growing burden of patient portal messaging on clinician time (Martinez et al., 2024; Mandal et al., 2024), LLM-powered QA workflows offer great promise for assisting clinicians through automatically generating EHR-grounded answers. Yet most efforts focus on patient question generation (Liu et al., 2024), with limited work on answer generation and existing studies of deployments relying on shallow, category-triggered EHR context with narrow evaluation of clinical adoption (Garcia et al., 2024; Bootsma-Robroeks et al., 2025). The ArchEHR-QA shared task aims to bridge this gap through the development of benchmarks and systems for holistic patient QA, spanning from question interpretation to evidence-grounded answer generation over free-form EHR notes (Soni et al., 2025; Soni and Demner-Fushman, 2026).

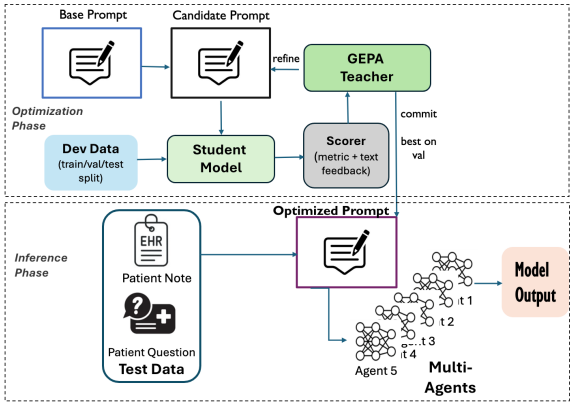

Figure 1: Overall Approach Diagram.

While fine-tuning LLMs on EHR records offers a promising avenue, its usability is hindered by high computational costs, task diversity and the scarcity of quality clinical annotations. Prompting frameworks have emerged as a superior alternative, surpassing fine-tuned LLMs in both general Medical QA (Maharjan et al., 2024) and Clinical QA tasks (Ray et al., 2026). Yet the effectiveness of prompting is contingent on prompt quality. Prior work has explored prompt engineering including

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 539–550

few-shot and Chain of Thought (CoT) for Clinical NLP tasks (Sivarajkumar et al., 2024), yet despite impressive zero-shot performance, such prompts remain task-specific, hand-crafted and unreliable across diverse question types (Zhu et al., 2024). Such ad-hoc approaches lack scalability and generalizability for evidence retrieval (Karayanni et al., 2024), while recent work has shown that jointly optimizing instructions and few-shot demonstrations outperforms standard prompting on Clinical EHR QA (Bogireddy et al., 2025).

In this work, we use GEPA (Agrawal et al., 2025), an evolutionary prompt optimization framework, combining it with multi-agent majority-voting enabling faithful and well-grounded question interpretation and answer synthesis without manual prompt engineering. We demonstrate our approach in Figure 1.

2. Related Work

Patient Question Interpretation has been studied as consumer health question summarization (Ben Abacha and Demner-Fushman, 2019; Basu et al., 2025), with work exploring transfer (Yadav et al., 2021b) and reinforcement learning (Yadav et al., 2021a). Element-aware generation, where key medical entities are extracted before synthesis, reduces hallucinations and achieves the highest entailment without few-shot prompting (Basu et al., 2025), while larger LLMs in zero-shot settings surpass fine-tuned models and show lower error rate than experts (Van Veen et al., 2023). Our approach adopts an entity-aware, template-driven abstraction strategy directly motivated by these findings.

Clinical EHR QA: Despite being explored across structured and unstructured data, Clinical EHR QA still remains underexplored and challenging (Bardhan et al., 2024). While emrQA (Pampari et al., 2018) established QA grounded on clinical EHRs, EHRNoteQA (Kweon et al., 2024) extended this to real longitudinal notes. AlMannaa et al. (2025) show performance benefits of hybrid RAG pipelines for retrieving relevant EHR sections in open-form QA, while highlighting challenges with long context and longitudinal reasoning. McInerney et al. (2023) and Ahsan et al. (2024) explored zero-shot prompting for EHR feature extraction and evidence retrieval respectively, flagging hallucinations. With real-world deployments remaining shallow and lacking grounded evidence retrieval (Bootsma-Robroeks et al., 2025), ArchEHR-QA (Soni and Demner-Fushman, 2026) directly addresses evidence identification and alignment over clinical notes.

Prompt Optimization demonstrates consistent gains over manual designs on QA tasks (Yang et al., 2023; Guo et al., 2023). Karayanni et al. (2024) show that expert-guided refinement outperforms both manual and automated optimization for clinical note classification, while Bogireddy et al. (2025) show that jointly optimizing instructions and demonstrations via MIPROv2 substantially outperforms zero-shot and few-shot baselines on grounded clinical note QA. Unlike MIPROv2's Bayesian search over fixed proposals, GEPA samples task trajectories and reflects on them to diagnose prompt failures and evolve candidates, significantly outperforming MIPROv2 (Agrawal et al., 2025). Here we employ GEPA alongside a multi-agent framework.

Multi-agent Voting: Multi-agent frameworks including debating reduce hallucination and improve performance (Du et al., 2024; Mao et al., 2025), with majority voting identified as the primary driver of these gains (Choi et al., 2025). In medical settings, collaborative multi-agent frameworks show strong zero-shot performance (Kim et al., 2024; Tang et al., 2024), with Hwang et al. (2025) achieving top performance on grounded EHR QA via sequential agent collaboration with dynamic expert recruitment, and majority voting outperforming debating in clinical QA abstract screening (Akinseloyin et al., 2026). Building on these findings, we employ majority voting across five diverse LLMs to leverage complementary model strengths.

3. Task Description

The ArchEHR-QA shared task addresses patient health question answering grounded in EHRs, spanning four subtasks:

T1 – Question Interpretation: Given patient question q, generate a concise clinician-interpreted query $\hat{q}'$ capturing the clinical rationale without introducing new facts.

T2 – Evidence Identification: Given q, clinician-interpreted query q' and the numbered sentences from the clinical note excerpt, $\{s_1, \ldots, s_n\}$, identify the *minimal sufficient set* of sentences $\hat{Y}^+ = \{i \mid \hat{y}_i = \texttt{essential}\}$ to answer q, remaining faithful to the excerpt even when it only partially addresses the question.

T3 – Answer Generation. Given q, $\hat{q}'$ and $\{s_1, \ldots, s_n\}$, generate a concise grounded answer directly addressing q, without speculation or use of external knowledge.

T4 – Evidence Alignment. Given q, $\hat{q}'$, $\{s_1, \ldots, s_n\}$ and answer $a = \{a_{s_1}, \ldots, a_{s_m}\}$, produce a many-to-many alignment mapping each answer sentence a_{s_j} to the note sentences that directly support it, yielding predicted pairs $[a_{s_j}, s_i]$.

3.1. Dataset and Evaluation

Data: The ArchEHR-QA 2026 dataset (Soni and Demner-Fushman, 2026) pairs patient-authored

questions q with clinical note excerpts from MIMIC (Johnson et al., 2016). Gold q' and a are provided where applicable (see Table 1). For Task 2 the sentence-level labels are: `essential` / `supplementary` / `non-relevant`.

Subtask	Dev	Test	Provided Inputs
T1: Patient Qn.	120	47	q
T2: Evidence	20	47	q, q', $\{s_i\}_{i=1}^n$
T3: Answer	20	47	q, q', $\{s_i\}_{i=1}^n$
T4: Alignment	20	147	q, q', $\{s_i\}_{i=1}^n$, a

Table 1: Dataset splits and inputs per subtask.

Evaluation: Each subtask is evaluated with automatic metrics as shown in Table 2. For Task 2, *strict* (only `essential`) and *lenient* (not penalizing `supplementary`) F1 variants are reported.

Metric	T1	T2	T3	T4
P / R / F1		•		•
BLEU (Papineni et al., 2002)			•	
ROUGE (Lin, 2004)	•		•	
SARI (Xu et al., 2016)			•	
BERTScore (Zhang et al., 2019)	•		•	
AlignScore (Zha et al., 2023)	•		•	
MEDCON (Yim et al., 2023)	•		•	

Table 2: Evaluation metrics per subtask.

4. Methodology

Our approach implements a systematic multi-stage pipeline using DSPy (Khattab et al., 2023), applied independently per subtask. We first iterate over multiple DSPy signatures, selecting the best strategy per task (§4.1). Following model selection across various LLMs under standardized clinical constraints (§4.2), we apply GEPA prompt optimization (§4.3) with task-specific textual feedback signals. For structured prediction (Tasks 2 and 4), we further employ multi-agent voting to improve reliability (§4.4).

4.1. DSPy Signature Iteration

We define multiple DSPy signatures per subtask, iterating over single and multi-step modules across different task-solving approaches. We experiment with two prompting strategies, vanilla zero-shot and chain-of-thought (CoT), holding the base signature fixed during GEPA optimization.

4.2. Model and Strategy Selection

We evaluate several state-of-the-art LLMs optimized for reasoning and zero-shot capabilities, comparing performance across standard inference and native reasoning modes. This ensures a standardized comparison of each model's inherent ability to adhere to clinical constraints, helping us identify the best candidate for evidence-grounding and answer generation over EHRs.

4.3. GEPA Prompt Optimization

After establishing base performance, we apply GEPA (Agrawal et al., 2025) within DSPy to automatically optimize module instructions. GEPA operates as a teacher–student loop where a teacher model samples trajectories from the student model to analyze failures and propose improved instructions. Formally, GEPA searches for the optimal prompt P^*:

$$P^* = \arg\max_{P \in \mathcal{P}} \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y^*) \in \mathcal{D}_{\text{train}}} \mathcal{R}\left(f_\theta(x;\, P),\, y^*\right) \tag{1}$$

where $P \in \mathcal{P}$ is a candidate prompt, $f_\theta(\cdot;\, P)$ is the student LLM conditioned on P, (x, y^*) is a training pair and $\mathcal{R}$ is the task-specific feedback reward.

Optimization targets two components sequentially:
(1) **instruction tuning**, where the teacher diagnoses student errors and rewrites instructions evolved from the Pareto frontier and
(2) **few-shot bootstrapping**, where high-quality demonstrations are automatically selected to accompany the optimized prompt.

We evaluate both stages during development. However, bootstrapped demonstrations did not overperform instruction-optimized zero-shot prompting, possibly due to the small number of examples available in dev set for demonstrations. Therefore all of our reported results reflect the instruction-optimized zero-shot prompting approach.

Feedback Signal: During initial evaluation (§4.1, 4.2) we use per-task metric averages (§3.1). During GEPA optimization, we additionally provide textual feedback alongside the numerical score to guide the teacher model when proposing new prompt candidates. For generative tasks (Tasks 1 and 3), an LLM judge (§A.2) generates a natural language explanation of how the student's output diverges from the gold reference. For structured prediction tasks (Tasks 2 and 4), we construct explicit feedback that identifies false positives (incorrectly predicted sentences) and false negatives (missed sentences) in the predicted evidence or alignment lists.

4.4. Multi-Agent Voting

For structured Tasks 2 and 4, we aggregate predictions from five LLMs via voting. After testing

Team	Avg	T1					T2	T3							T4
		O.	R	BS	AS	MED	S.F1	O	B	R	SARI	BS	AS	MED	F1
OptiMed(ours)	**52.0**	29.9	28.8	43.1	**27.7**	19.9	63.2	34.5	5.7	25.2	56.5	43.1	**37.4**	39.1	79.1
Neural	51.6	28.9	31.3	43.6	15.2	25.6	**63.7**	35.2	9.4	25.6	57.7	43.5	34.3	40.9	80.3
WisPerMed	50.8	26.9	21.8	38.0	21.7	**26.3**	58.8	**36.3**	**9.9**	**27.8**	58.6	**46.8**	31.7	43.1	**81.5**
HealthNLP_R.	50.7	**31.2**	**35.3**	**46.8**	24.0	18.7	60.2	34.6	7.0	25.4	59.2	43.8	33.6	38.7	79.8
Yale-DM-Lab	50.1	27.1	28.2	40.6	19.7	19.8	61.9	31.0	9.1	23.1	56.5	37.2	22.1	37.6	67.2
BIT.UA-AAUBS	48.7	19.0	17.9	29.4	8.7	20.2	58.8	35.6	8.6	26.4	**60.0**	45.0	30.2	**43.2**	80.4
sebis	45.9	25.6	22.9	36.9	21.4	21.3	51.6	31.5	3.8	22.0	56.6	39.3	33.9	33.4	74.8

Table 3: Results on the ArchEHR-QA 2026 test set for teams completing all four subtasks, ranked by average overall score. O. = Overall, B = BLEU, R = ROUGELsum, BS = BERTScore, AS = AlignScore, MED = MEDCON (UMLS), S.F1 = Strict Micro F1, F1 = Micro F1.

different voting thresholds, we find strict majority ($\geq 3/5$) to perform best:

$$\hat{y}_i = \mathbf{1}\left[\sum_{k=1}^{5} \hat{y}_i^{(k)} \geq 3\right] \tag{2}$$

improving reliability and recall while suppressing single-model errors. The shared agent instruction was optimized via GEPA. We further investigate the effect of upstream evidence quality on Tasks 3 and 4 by using the multi-agent ensemble evidence from Task 2 as input in a two-step approach.

5. Experimental Setup

All models are accessed via OpenRouter with no fine-tuning.

GEPA Teacher: `gemini-3.0-flash`, selected for its performance, context size, and efficiency as the optimizer.

Single-model Inference: `glm-4.7`, with strong instruction-following and cost efficiency.

Multi-agent Voting: Ensemble of the five following zero-shot LLMs: `deepseek-v3.2`, `glm-4.7`, `gemini-3.0-flash`, `gpt-5.2`, and `kimi-k2.5`, maximizing diversity across training data, architecture, and provider, identified as the key driver of majority voting gains (Akinseloyin et al., 2026).

GEPA Prompt Optimization Setup: For GEPA optimization, the development data is partitioned into train, validation and held-out test subsets: 80/20/20 for Task 1 (120 examples) and 7/3/10 for Tasks 2, 3, and 4 (20 examples). GEPA iterates over the training subset, using textual feedback to propose new prompt candidates, committing each only if it improves performance on the validation subset. The held-out test subset is used for selecting the best-performing overall system configuration across models and approaches prior to final evaluation on the shared task test set.

Inference Parameters: All single-model runs use greedy decoding (temperature $= 0$), while ensemble runs use temperature $= 0.3$ to promote output diversity. A maximum of 32,000 output tokens is used throughout. Where supported, native reasoning was disabled, as enabling it yielded worse performance on the dev set.

6. Results

Table 3 shows performance across teams completing all four subtasks. OptiMed ranked **1st** overall among 7 full-task teams with a score of **52.0**, and among all 23 teams ranked 3rd in T1 (**29.9**), 2nd in T2 (**63.2**), 6th in T3 (**34.5**), and 4th in T4 (**80.3**). Notably, we achieve the top AlignScore across all teams in both Tasks 1 and 3, reflecting superior factual grounding relative to the ground truth, avoiding hallucination which is particularly meaningful in medical settings. These results demonstrate that GEPA optimization and multi-agent majority voting complement each other across tasks, forming a strong candidate for clinical settings, though GEPA's gains are contingent on sufficient dev set size for reliable prompt optimization. We discuss task-specific findings below.

Model	Task 1	Task 2	Task 3
`glm-4.7`	**35.7**	**50.1**	37.3
`gemini-3.0-flash`	26.5	37.5	**40.5**
`kimi-k2.5`	–	**50.1**	37.1
`gpt-5.2`	21.6	49.1	38.4
`sonnet-4.5`	23.9	37.3	–

Table 4: Model performance on dev set.

Model Selection: Table 4 shows individual model performance on the dev set (see Table 1). `glm-4.7` achieves the highest scores on Tasks 1 and 2, making it a strong choice for a single base model on these tasks. On Task 1, it specifically surpasses all other models by a very large margin. However, on Task 3 `gemini-3.0-flash` has the best performance, motivating its selection as the base model for answer generation using GEPA, with `glm-4.7` retained as a strong candidate for its consistently high performance across tasks.

System	O.	R	BS	AS	MED
Base prompt + CoT	27.8	26.0	40.5	22.6	**22.0**
GEPA	29.3	28.5	42.8	23.6	22.3
GEPA + CoT	**29.9**	**28.8**	**43.1**	**27.7**	19.9

Table 5: Zero-shot performance for Task 1 with `glm-4.7` model on test set.

Question Interpretation: Motivated by the superiority of element-aware generation (Basu et al., 2025), our base prompt identifies the patient's concern, extracts relevant medical entities (symptoms, events and therapies) and maps them to a standard query template based on the question type (treatment, cause, plan questions, care questions) with emphasis on clinical formalization. Clinical formalization refers to the structured reframing of informal patient language into a concise, clinician-facing query (see Appendix A.1). Table 6 shows results for three systems using `glm-4.7`, where GEPA yields a 1.5-point gain over the base prompt, with CoT providing further improvements.

System	S.F1	S.P	S.R	L.F1	L.P	L.R
GEPA w. q, q'	59.4	**57.8**	61.0	67.1	**74.6**	61.0
GEPA w. q	61.9	52.0	**76.6**	71.0	66.2	**76.6**
GEPA w. q + Multi-agent (3)	**63.2**	56.7	71.3	**72.2**	73.1	71.3

Table 6: Zero-shot performance on Task 2 with `glm-4.7` on test set on micro F1. S. = Strict, L. = Lenient.

Evidence Identification: With `glm-4.7` remaining the best on the dev set, we compare two GEPA variants: providing q only versus both q and q'. The latter yields higher Precision but lower Recall and overall lower F1, showing that q''s conciseness focuses evidence selection more narrowly. Multi-agent majority voting (3/5) across five diverse LLMs achieves the best overall performance.

System	O.	B	R	SARI	BS	AS	MED
GEPA (Gemini) Multi-Agent from T2 (thr=2)	30.0	4.2	23.4	54.6	41.5	23.8	32.5
GEPA (Gemini) Multi-Agent from T2 (thr=3)	31.1	5.2	23.9	55.0	42.1	26.3	33.9
Base prompt (GLM) Only GLM from T2	**34.5**	**5.7**	**25.2**	**56.5**	**43.1**	**37.4**	**39.1**

Table 7: Zero-shot performance on Task 3 test set. T2 denotes using the Task 2 evidence output as input in a *two-step* approach.

Answer Generation: We adopt a two-stage approach, using Task 2 evidence as input for Task 3. For generation, the simpler base prompt using `glm-4.7` significantly outperforms the GEPA variant using `gemini-3.0-flash` for both the majority (3 out of 5 LLMs) and lenient (2 out of 5 LLMs) multi-agent voting variants. Additionally, the majority (strict) variant of voting surpasses the lenient across all the metrics. We attribute GEPA's failure to the task complexity and diversity in combination with the small dev set, which provides insufficient signal to reliably distinguish prompt quality from instance-level variance.

System	Micro F1	Micro P	Micro R
Base prompt + CoT Multi-Agent from T2 (thr=3)	62.6	64.3	61.0
Base prompt + Multi-agent (3)	**80.3**	80.7	**79.8**
GEPA + Multi-agent (3)	79.3	**80.5**	78.1

Table 8: Zero-shot performance on Task 4 with `glm-4.7` model on test set. T2 denotes using the Task 2 evidence output as input in a *two-step* approach, while Multi-agent refers to a *single-step* multi-agent approach without Task 2 input.

Evidence Alignment: Based on our Task 2 findings, we omit q'. A two-step approach using Task 2 evidence identification followed by `glm-4.7` CoT alignment yielded poor performance, likely due to error propagation and single-model high uncertainty. Switching to single-step multi-agent majority voting that does not use Task 2 as input improved results, with the base variant outperforming GEPA, consistent with our Task 3 finding that GEPA fails to generalize on small dev sets and complex tasks such as many-to-many evidence alignment.

6.1. GEPA-Optimized Prompts Insights

Base and GEPA-optimized prompts per task are provided in Appendix A.1. For Tasks 1 and 2, where GEPA outperformed the base prompt, gains appear driven by role prompting that places the model in a clinical context (i.e. *You are a reliable expert clinician*) and fine-grained procedural instructions that decompose the task into verifiable steps. For Tasks 3 and 4, the GEPA prompts are markedly more complex, introducing formal auditing frameworks and domain-specific sub-rules. We hypothesize that the combination of inherently more complex tasks with very limited training examples produces high variance in the optimization signal, causing the teacher model to over-specify instructions rather than learning robust task-level patterns, resulting in prompts that do not generalize well.

7. Conclusion

We present OptiMed, combining GEPA prompt optimization and multi-agent majority voting for grounded clinical EHR question answering, ranking 1st overall among teams completing all four subtasks and achieving the top AlignScore in generative tasks. GEPA proved effective where training signal was sufficient, while multi-agent voting yielded notable gains on complex structured prediction tasks. For complex tasks with small dev sets, GEPA underperformed base prompts, highlighting the importance of sufficient supervision signal.

8. Limitations

While GEPA optimization shows benefits, it is sensitive to development set size, conditioning its success on the quality and quantity of clinical annotations, especially for complex generative tasks. The real performance gains under greater data availability therefore remain to be examined. Regarding cost, multi-agent voting across five LLMs amplifies inference costs compared to single-LLM counterparts, which may hinder deployment in real clinical settings. Our system is evaluated on a single benchmark dataset, and thus generalization to other EHR systems with different structures and more diverse patient populations remains untested. Additionally, the effectiveness of such a system for patients from diverse racial and ethnic backgrounds remains unknown, as their conditions and clinical reasoning patterns may be underrepresented in LLM training data, potentially compromising both accuracy and fairness. Finally, while our system achieves strong automatic metric performance, the clinical validity of generated answers and evidence alignments has not been verified through human expert evaluation which is essential before any real-world deployment.

9. Ethics

This work uses the ArchEHR-QA 2026 dataset, which is derived from MIMIC (Johnson et al., 2016) and accessed under the PhysioNet Credentialed Health Data License `1.5.0`. The dataset contains de-identified patient records in accordance with HIPAA Safe Harbor standards. All experiments were conducted solely on this de-identified data, and no attempts were made to re-identify any individuals.

10. Acknowledgements

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

11. Bibliographical References

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. 2025. Gepa: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*.

Hiba Ahsan, Denis Jered McInerney, Jisoo Kim, Christopher Potter, Geoffrey Young, Silvio Amir, and Byron C Wallace. 2024. Retrieving evidence from ehrs with llms: possibilities and challenges. *Proceedings of machine learning research*, 248:489.

Opeoluwa Akinseloyin, Xiaorui Jiang, and Vasile Palade. 2026. Llm-based multi-agent collaboration for abstract screening towards automated systematic reviews. *Biology Methods and Protocols*, page bpag006.

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

Feras AlMannaa, Talia Tseriotou, Jenny Chim, and Maria Liakata. 2025. Investigating LLM capabilities on long context comprehension for clinical question answering. *arXiv preprint arXiv:2510.18691*.

Jayetri Bardhan, Kirk Roberts, and Daisy Zhe Wang. 2024. Question answering for electronic health records: Scoping review of datasets and models. *J Med Internet Res*, 26:e53636.

Abhishek Basu, Deepak Gupta, Dina Demner-Fushman, and Shweta Yadav. 2025. A dataset and benchmark for consumer healthcare question summarization. *arXiv preprint arXiv:2512.23637*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Sai Prasanna Teja Reddy Bogireddy, Abrar Majeedi, Viswanath Gajjala, Zhuoyan Xu, Siddhant Rai, and Vaishnav Potlapalli. 2025. Neural at ArchEHR-QA 2025: Agentic prompt optimization for evidence-grounded clinical question answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 104–109, Vienna, Austria. Association for Computational Linguistics.

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

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. 2025. Debate or vote: Which yields better decisions

in multi-agent large language models? In *Advances in Neural Information Processing Systems*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2023. Evoprompt: Connecting llms with evolutionary algorithms yields powerful prompt optimizers. *arXiv e-prints*, pages arXiv–2309.

Hyeon Hwang, Hyeongsoon Hwang, Jongmyung Jung, Jaehoon Yun, Minju Song, Yein Park, Dain Kim, Taewhoo Lee, Jiwoong Sohn, Chanwoong Yoon, Sihyeon Park, Jiwoo Lee, Heechul Yang, and Jaewoo Kang. 2025. DMIS lab at ArchEHR-QA 2025: Evidence-grounded answer generation for EHR-based QA via a multi-agent framework. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 118–125, Vienna, Austria. Association for Computational Linguistics.

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, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: an adaptive collaboration of llms for medical decision-making. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

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.

Siru Liu, Aileen P Wright, Allison B Mccoy, Sean S Huang, Julian Z Genkins, Josh F Peterson, Yaa A Kumah-Crystal, William Martinez, Babatunde Carew, Dara Mize, Bryan Steitz, and Adam Wright. 2024. Using large language model to guide patients to create efficient and comprehensive clinical care message. *Journal of the American Medical Informatics Association*, 31(8):1665–1670.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3):100943.

Jenish Maharjan, Anurag Garikipati, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Rahul Thapa, Qingqing Mao, and Ritankar Das. 2024. Openmedlm: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, 14(1):14156.

Soumik Mandal, Batia M Wiesenfeld, Devin M Mann, Adam C Szerencsy, Eduardo Iturrate, and Oded Nov. 2024. Quantifying the impact of telemedicine and patient medical advice request messages on physicians' work-outside-work. *NPJ Digital Medicine*, 7(1):35.

Junyu Mao, Anthony Hills, Talia Tseriotou, Maria Liakata, Aya Shamir, Dan Sayda, Dana Atzil-Slonim, Natalie Djohari, Arpan Mandal, Silke Roth, et al. 2025. Automated data enrichment using confidence-aware fine-grained debate among open-source LLMs for mental health and online safety. *arXiv preprint arXiv:2512.06227*.

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

Denis McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron Wallace. 2023. CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8477–8494, Singapore. Association for Computational Linguistics.

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*.

Jie Pan, Seungwon Lee, Cheligeer Cheligeer, Elliot A. Martin, Kiarash Riazi, Hude Quan, and Na Li. 2025. Integrating large language models with human expertise for disease detection in electronic health records. *Computers in Biology and Medicine*, 191:110161.

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.

Sushant Kumar Ray, Gautam Siddharth Kashyap, Sahil Tripathi, Nipun Joshi, Vijay Govindarajan, Rafiq Ali, Jiechao Gao, and Usman Naseem. 2026. Do clinical question answering systems really need specialised medical fine tuning? *arXiv preprint arXiv:2601.12812*.

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*, 31(3):943–950.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Med Inform*, 12:e55318.

Sarvesh Soni and Dina Demner-Fushman. 2026. Overview of the archehr-qa 2026 shared task on grounded question answering from electronic health records. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma, Mallorca (Spain). ELRA.

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

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

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. 2023. Clinical text summarization: adapting large language models can outperform human experts. *Research square*, pages rs–3.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. volume 4, pages 401–415.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2021a. Reinforcement learning for abstractive question summarization with question-aware semantic rewards. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 249–255, Online. Association for Computational Linguistics.

Shweta Yadav, Mourad Sarrouti, and Deepak Gupta. 2021b. NLM at MEDIQA 2021: Transfer learning-based approaches for consumer question and multi-answer summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 291–301, Online. Association for Computational Linguistics.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

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. *arXiv preprint arXiv:1904.09675*.

Yinghao Zhu, Zixiang Wang, Junyi Gao, Yuning Tong, Jingkun An, Weibin Liao, Ewen M Harrison, Liantao Ma, and Chengwei Pan. 2024. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint arXiv:2402.01713*.

12. Language Resource References

Jayetri Bardhan, Anthony Colas, Kirk Roberts, and Daisy Zhe Wang. 2022. DrugEHRQA: A question answering dataset on structured and unstructured electronic health records for medicine related queries. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1083–1097, Marseille, France. European Language Resources Association.

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

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

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Sarvesh Soni and Dina Demner-Fushman. 2026. A dataset for addressing patient's information needs related to clinical course of hospitalization. *Scientific Data*.

A. Appendix

A.1. Base versus GEPA-Optimized Prompts

For each Task we show the hand-crafted **Base** prompt alongside the **GEPA-optimized** prompt.

Task 1 – Question Interpretation

Base Prompt – Task 1

Analyze a patient's story to figure out the patient's main concern/question and formulate the single question a clinician would use to look up in the EHR record.

Task:

1. Locate the specific *Conflict*, *Gap of information*, or *Confusion*:
   - Is it about a Treatment Choice?
   - Is it about a Cause?
   - Is it about the Future/Plan?
   - Is it about Quality of Care?
2. Identify the main *Event/Condition*, *Symptom*, and *Therapy*:
   - Are any of these directly related to the patient's question?
   - If yes, only then include them in the output question.
3. Map to a Standard Clinical Query Template:
   - Treatment → "Why was [Therapy] recommended/given?"
   - Cause → "Is [Symptom] related to [Event/Condition]?"
   - Plan → "What is the expected course of recovery?" or "What should the patient do?"
   - Care → "Why was [Diagnosis/Treatment] delayed?" or "Why was [Decision] made?"

GEPA-Optimized Prompt – Task 1

Act as a clinical intake specialist. Your task is to extract the core medical concern from a patient's narrative and rephrase it into a single, professional *Clinician Question*.

Follow these steps:

1. **Identify the Medical Catalyst:** Locate primary symptoms (e.g., "numbness," "chest pain"), specific events (e.g., "rat bite," "surgery"), or medications mentioned.
2. **Filter the "Noise":** Remove all emotional context, pleasantries, expressions of gratitude, and conversational fillers.
3. **Synthesise the Query:** Formulate the question focused on the clinical objective—seeking a diagnosis (*What caused X?*), causality (*Was X caused by Y?*), or triage (*Should I take action?*).
4. **Maintain Terminology:** Use the patient's specific language for symptoms and medications but organise it with professional structure.

5. **Standardise the Phrasing:**
    - Use simple starters: "What caused…", "Was the…", "Should I…", "Is it expected…", "Why was…".
    - Convert vague worries like "What happened?" into "What did the findings show?" or "What caused these symptoms?".
    - Convert "Is that normal?" to "Is it expected for [symptom] to occur?".
    - Ensure the final output is a single, concise sentence a clinician can address immediately.

Task 2 – Evidence Identification

Base Prompt – Task 2

Given a patient's medical question and a numbered set of sentences from a clinical note excerpt, identify the sentences that are essential to answering the question.

A sentence is *essential* if removing it would leave the answer to the patient's question incomplete — it directly provides clinical evidence (findings, procedures, results, reasoning, or recommendations) that addresses what the patient asked.

The goal is precision: include every sentence without which the answer would be incomplete, but exclude sentences that merely provide context or are unrelated.

GEPA-Optimized Prompt – Task 2

You are a reliable expert clinician. Given a patient's medical narrative and a numbered set of sentences from a clinical note excerpt, identify all sentence IDs that are essential to answering the question.

Definition of "Essential": A sentence is essential if its removal results in an incomplete answer. It must provide direct clinical evidence, including:

1. **Findings & Results:** Specific test results, diagnoses, or negative findings that rule out or confirm a condition.
2. **Clinical Reasoning:** The rationale behind a diagnosis or treatment path, including evidence of clinical deterioration that justified a procedure.
3. **Recommendations & Action Plans:** Specific follow-up instructions, including doctor names or specialties for further care.
4. **Safety Information (Red Flags):** Warning signs the patient is told to monitor and report immediately.

Exclusion Criteria: Exclude sentences providing only general context, hospital pleasantries, technical procedural details without outcome relevance, or redundant information.

Strategy by Question Type:

- *Recovery / "What do I do?"*: Include prognosis, follow-up plan, and red flags.
- *"Is it connected?"*: Include positive evidence and significant negatives used to rule out connections.
- *"Why this procedure?"*: Focus on failure of prior treatments, worsening lab trends, and obstructions found.

Output: Provide only a list of the essential sentence IDs.

Task 3 – Answer Generation

Base Prompt – Task 3

You are a medical expert tasked with answering a patient's question using only information explicitly stated in the provided clinical note excerpt.

Task: Using only the essential sentences identified by the provided indices, write a response that directly and faithfully addresses the patient's question.

Rules:

- Use only facts explicitly stated in the clinical note. Do not speculate or infer beyond what is documented.
- If the note does not fully answer the question, answer what can be supported and acknowledge the gap. Do not fabricate information.
- Write in clear, simple language a patient or family member can understand.
- Do not include citations, sentence numbers, or references.
- Write a 3rd-person, concise, plain prose response directly addressing the patient's question.
- Keep the answer under 75 words (≈5 sentences).

GEPA-Optimized Prompt – Task 3

You are a medical expert assistant. Extract and translate key clinical information from a patient's medical record into a concise, medically precise explanation for the patient or their family.

Task Instructions:

1. **Strict Evidence Filtering:** Use *only* the sentences corresponding to the provided `evidence_ids`.
2. **Focus on Clinical Reasoning:** Explain *why* medical actions were taken by connecting phys-

ical exam findings and diagnostic results to treatments.

3. **Symptom Documentation:** Incorporate all symptoms and physical exam findings mentioned in the filtered evidence.
4. **Synthesise Findings:** Do not include specific lab values or exact dosages; use interpretations instead (e.g., "elevated muscle enzymes" rather than "CK peaking at 1405").
5. **Closing the Loop:** Directly address the patient's concern with an evidence-based answer.

Strict Rules:

- **Factuality:** Do not use outside medical knowledge.
- **Tone & Perspective:** Write in the 3rd person. Use simple, clear sentences while retaining professional terminology.
- **Format:** Max 75 words; no citations or sentence numbers; single seamless paragraph.
- **Goal of Care:** Explicitly mention any transition to comfort measures or major shift in the direction of care if evidenced.

Task 4 – Evidence Alignment

Base Prompt – Task 4

Ground each sentence of a clinical answer to the specific sentence(s) in the clinical note excerpt that directly support it.

Alignment Rules:

- Operate at the answer-sentence level: each answer sentence maps to zero, one, or more note sentences.
- Only link a note sentence when it provides direct, specific support — not merely topical overlap.
- Alignments are many-to-many: one note sentence may support multiple answer sentences and vice versa.
- Answer sentences derived entirely from clinical knowledge outside the excerpt map to an empty list.
- Cite only note sentences that genuinely ground the claim in the answer sentence.

GEPA-Optimized Prompt – Task 4

You are a clinical data analyst performing a high-fidelity audit of patient-facing explanations. Map simplified *Answer Sentences* to original *Note Sentences* using the standard of **Minimal Sufficient Proof**.

Core Principles:

1. **Minimal Sufficient Proof:** Select the smallest set of note sentences that together provide 100% factual support. Apply the *Redundancy Test*: if removing a note sentence leaves the claim fully proven, remove it.
2. **Strict Claim Decomposition:** Break each answer sentence into atomic claims. Every claim must have an explicit corresponding fact in the note; if any claim is missing or only implied, the alignment for that sentence is `[]`.
3. **Cross-Sectional Mapping:** If an answer sentence uses both technical and layman terms, select both the technical evidence and the summary evidence — but do not include underlying data when a single discharge instruction sentence already covers the claim.

Domain-Specific Rules:

- *Cardiac rule-outs:* Require both diagnostic data (troponins/EKGs) and the physician's conclusion.
- *Musculoskeletal pain:* Require physical exam findings and the clinical conclusion.
- *Follow-up instructions:* Must identify doctor/specialty, timeframe/test, and purpose; incomplete matches align to `[]`.
- *Headers:* Include only if the header is the sole source of the specific diagnosis in the answer sentence.

Output Format: Precede alignments with a **Reasoning** section listing atomic claims and justifying the minimal evidence set, followed by a JSON-style dictionary: `{"1": [i, j], "2": [], "3": [k]}`.

A.2. LLM Judge Prompts for GEPA Optimization

For generative tasks (Tasks 1 and 3), GEPA uses an LLM judge to generate concise textual feedback by comparing the student's output against the gold reference, guiding the teacher model in proposing improved prompt candidates.

Judge Prompt — Task 1: Q. Interpretation

You are an expert Medical QA evaluator. You are evaluating a system's ability to transform a free-text, patient-authored narrative into a clear and concise clinician-interpreted question. The goal is to maximize the similarity between the predicted clinical question and the gold question.

Feedback guidelines:

- Should be general and not focused on this sample.
- Should be concise and focused only on why the prediction differs from the gold.
- Should guide the system on how to improve its performance on this task in general.
- Do not include samples from the provided narrative.
- Keep it under 20 words.

Judge Prompt — Task 3: Answer Generation

You are an expert Medical QA evaluator. You are evaluating a system's ability to answer a patient-authored question using only information explicitly stated in the provided clinical note excerpt. The goal is to maximize the similarity between the predicted answer and the gold answer.

Feedback guidelines:

- Should be general and not focused on this sample.
- Should be concise and focused only on why the prediction differs from the gold.
- Should guide the system on how to improve its performance on this task in general.
- Do not include samples from the provided narrative.
- Keep it under 20 words.