MedEvi-NS at ArchEHR-QA 2026: Using Clinical Reasoning Principles to Improve Zero-shot Capabilities of Large Language Models in Evidence Alignment

Mengxuan Sun[1] *, **Nicolay Rusnachenko**[2] *
[1] Institute of Applied Health Sciences, University of Aberdeen, UK,
[2] Centre for Applied Creative Technologies (CFACT+), Bournemouth University, UK
m.sun.22@abdn.ac.uk, nrusnachenko@bournemouth.ac.uk
* *These authors contributed equally to this work.*

Abstract

The ArchEHR-QA shared task focuses on grounded question answering using patient EHR data. For the given clinical interpretation of the patient question, note excerpt (E) and answer text (A), subtask 4 (evidence alignment) aims to cite supporting sentences from E for each sentence in A. In this paper, we propose a prompt-engineering methodology that features clinical-reasoning principles in related alignment. We adopt this methodology for GPT-5.2 in zero-shot learning mode. According to our experiments on ArchEHR-QA, incorporating clinical reasoning principles into the prompt improves $F1_{overall}$ by +2.0%. Our final submission resulted in 77.4% by $F1_{overall}$, which positions us at 10th out of 16 teams. Our code is publicly available: https://github.com/nicolay-r/ArchEHR-QA-2026-Task-4-MedEvi-NS

Keywords: Large Language Models, Evidence Alignment, Prompt-based Engineering, Clinical Reasoning

1. Introduction

Recent advances in large language models (LLMs) have significantly improved question answering systems across many domains (Shailendra et al., 2024). However, applying such systems to clinical use cases remains challenging as hallucinated or unsupported responses may pose patient safety risks (Kim et al., 2025). In clinical practice, evidence-based medicine emphasises that medical conclusions should be supported by explicit clinical evidence (Lohr et al., 1998). Therefore, ensuring accurate alignment of evidence is essential for maintaining the faithfulness in clinical question answering.

The ArchEHR-QA shared task focuses on grounded question answering using patient EHR data (Soni and Demner-Fushman, 2026a,b). The objective is to answer patient questions and provide evidence supported by clinical notes. In this work, we participate in subtask 4, which requires aligning each answer sentence with the specific supporting sentence(s) in the clinical note excerpt.

Existing approaches to evidence alignment are often formulated as evidence retrieval or semantic matching tasks (Zhao et al., 2024; Karthik et al., 2025; Gupta et al., 2024). However, such approaches may be insufficient for clinical narratives. Clinical records frequently describe multiple medical events and treatment stages within the same document (Hazlehurst et al., 2005). As a result, surface-level semantic matching may retrieve sentences that share keywords or embeddings with the answer but refer to different clinical episodes or stages of care.

1. **Clinical episode identification**
   locate the medical event referenced in the answer. (e.g., surgery, examination, or hospitalisation).
2. **Stage of care determination**
   determine the stage of the treatment process the answer describes.
   background → finding → decision → procedure → outcome.
3. **Evidence retrieval**
   search the clinical note for sentences describing the same episode and stage.
4. **Evidence verification**
   ensure that each clinical fact in the answer is supported by at least one cited sentence.

Figure 1: Four-step clinical reasoning principles declaration, utilised as a part of our zero-shot learning methodology.

Recent studies have explored the use of large language models (LLMs) and prompting techniques to support clinical tasks. For example, prompt engineering guidelines have been proposed to help clinicians design effective prompts by defining clinical objectives, following medical reasoning principles, and evaluating prompt quality (Liu et al., 2025). Other work investigates instruction tun-

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 482–486
12 May 2026.

ing and prompt optimisation strategies to improve LLM performance in medical applications (Le et al., 2025; Dhamija and Sharma, 2025).

Motivated by these developments, we propose a prompt-based evidence-alignment method that leverages large language models and incorporates principles of clinical reasoning, using GPT 5.2 by OpenAI. Instead of relying on surface-level similarity, our approach guides the model to identify the clinical episode and stage of care referenced in each answer sentence. The clinical reasoning principles are shown in Figure 1. Experimental results on ArchEHR-QA show that incorporating structured clinical reasoning improves the evidence alignment performance.

2. Methodology

Task Definition:[1] Given a case (Q_c, E, A), where Q_c is clinician-interpreted question, $E = \{e_1, e_2, \ldots, e_m\}$ is a *clinical note excerpt* that consist of sentences $e_i, i \in \overline{1,m}$, and $A = \{a_1, a_2, \ldots, a_n\}$ is an *answer* that yields of sentences $a_i, i \in \overline{1,n}$. For each $a_i \in A$, the task is to identify supporting sentences (E_s), where $E_s \subseteq E$.

Our methodology represents prompt-based engineering, with the structure of the *clinical prompt* depicted in Figure 2, with the details on the prompt components covered in Figure 3. It features **clinical reasoning principles** which are in greater detail covered in Section 2.1.

Clinical Prompt

You are grounding answer sentences to a clinical note.

Task Declaration

Clinical reasoning principles

Output Format

Input Description

Output (each answer sentence followed by its supporting sentence number(s) in brackets):

Figure 2: Clinical Prompt which features clinical reasoning principles (see Section 2.1)

For the given input example (Q_c, E, A), we compose prompt to query the model to align the whole answer (A) in zero-shot[2] learning mode.

[1] Our methodology omits patient question.

[2] Our approach could be viewed as "one-shot" since *Output Format* includes a single-sentence example (Figure 3); however the intention for including the example is to propose the expected output format for the task

Task Declaration

Goal:
For each answer sentence, identify the clinical note sentence(s) that support it.

Input:
Clinician-interpreted question (Clinician-Interpreted Question)
Clinical note excerpt with numbered sentences (Clinical Note Excerpt)
Answer text to be grounded (Reference Answer)

Expected output:
For each answer sentence, a set of supporting evidence sentence numbers from the clinical note excerpt.

Clinical reasoning principles

Clinical documentation follows a progression:
background → finding → decision → procedure → outcome.
For each answer sentence:

List of Rules

If any clinical fact is unsupported, search again and add the necessary sentence. Do not output this verification step.

Output Format

Output format:
Output each answer sentence followed by its supporting sentence number(s) in brackets.
Example: "The patient had elevated bilirubin [3]. ERCP was performed [5]."

Input Description

- - -
Clinician-Interpreted Question:
Q_c
Clinical Note Excerpt (numbered sentences):
E
Reference Answer (ground each sentence to note sentences above):
A
- - -

Figure 3: Components of the clinical prompt and with content for "list of rules", described in Section 2.1; for the *Output Format*, we include a fixed "example" for leveraging in-context learning; bold is used for highlighting the key words that are used in the prompt.

Structuring the output. To structure the raw output, we applied the following steps: (i) identifying

list of sentences/chunks (ii) mapping the most likely relevant sentence for the chunk.

We use "\n" char to demarcate content into sentences. We apply regular expression to retrieve content in "[]" by covering: separated entries, dash-separated entries (ranges). We use the Levenshtein distance (Levenshtein et al., 1966) to determine the best match between the output chunk and the sentence.

2.1. Clinical Reasoning Principles Prompt Design

Clinical records typically document the progression of a patient's treatment, including clinical findings, treatment decisions, procedures performed, and outcomes. In evidence-based medicine, clinicians interpret treatment conclusions by associating clinical observations with specific medical events and stages of care (Zakowski et al., 2004).

Inspired by this process, we incorporate principles of clinical reasoning into prompt design to guide evidence alignment. Specifically, the model needs to reason according to the following steps:

Step 1: Clinical episode identification. The model first identifies the clinical episode referenced by the answer sentence, such as a specific surgery, examination, or hospitalisation. Since clinical terms could refer to multiple episodes, it is easy to misalign the answer with sentences describing other events if events are not distinguished. Therefore, we explicitly require the model in the prompt to only look for evidence within the scope of the corresponding clinical episode.

Step 2: Stage-of-care identification. After identifying the episode, the model needs to determine which stage of the treatment process the answer sentence describes. We categorise common treatment stages in clinical records into four types: clinical findings, treatment decisions or recommendations, procedures performed, and outcomes. The model needs to determine the stage to which the answer sentence belongs based on its semantics, then search for sentences in the clinical records that describe the same event and stage as evidence.

Step 3: Evidence selection rules. To ensure reliable grounding, we designed several rules to strengthen the principle and avoid misalignment.

- Sentences that clearly show that clinical events have happened should be given first priority. For example, statements that say surgery has been done or a diagnostic result has been seen.
- Avoid selecting sentences that only share keywords with the answer but talk about clinical events or background information that isn't relevant.

List of Rules

1. Identify which clinical episode it refers to (e.g., first ERCP, repeat ERCP, specific hospital day). Only search within that episode.
2. Determine what stage of care the answer sentence represents:
   - a clinical finding
   - a decision/recommendation
   - a performed procedure
   - a result/outcome
3. Select the sentence(s) in the clinical note that directly document that specific stage of care.
4. Prefer explicit documentation of completed events (e.g., "was performed", "showed", "was found") rather than related background or earlier mentions.
5. Do NOT cite sentences:
   - from a different episode
   - that only share keywords
   - that describe a different stage of care
6. Cite the minimal number of sentences needed. Only cite a sentence if removing it would make the answer unsupported.
7. Before finalising the citation set for an answer sentence, verify that every distinct clinical fact expressed in the sentence is explicitly supported by at least one cited clinical note sentence.

Figure 4: List of Rules that a part of clinical reasoning principles

- Select the minimum set of sentences required to support the answer statement.

Figure 4 lists evidence selection rules.

Step 4: Evidence Verification. Finally, we added a validation step that requires the model to verify that each clinical fact expressed in the answer sentence is supported by at least one citation from a clinical record, and to check the answer before giving the final answer.

3. Dataset

We use the officially provided ArchEHR-QA dataset (D) version 1.5 (Soni and Demner-Fushman, 2026a). Table 2 presents a verbose analysis of cases for excerpt and answer lengths, separately for development (D_{dev}) and test (D_{test}) splits. According to the related comparison, with nearly similar length of sentences in excerpts and answers, **answers are ≈ 5.32 times shorter** (in sentences) than excerpts.

Method	$F1_{overall}$	P_{micro}	R_{micro}	$F1_{micro}$	P_{macro}	R_{macro}	$F1_{macro}$
			D_{dev}				
LLaMA-3-70B-instruct (‡)	67.95	68.28	72.26	70.21	75.51	76.29	65.68
GPT-5.2 (†)	74.77	70.69	89.13	78.85	73.01	90.50	70.68
GPT-5.2 (‡)	76.24	77.48	84.78	80.97	79.39	86.87	71.51
			D_{test}				
GPT-5.2 (‡) (ours)	77.41	75.70	77.38	76.53	79.28	80.81	78.29
team `boris123` (baseline)	–	76.50	58.20	66.10	–	–	–

Table 1: Results on ArchEHR-QA separately for D_{dev} and D_{test} splits; † - excluding reasoning principles from the prompt, ‡ - including reasoning principles (Section 2.1).

Parameters	D_{dev}	D_{test}
Cases (Total)	20	147
Excerpt (chars / sentence avg)	95.98	93.13
Excerpt (sentences avg)	21.4	28.5
Excerpt (sentences min)	9	5
Excerpt (sentences max)	54	73
Answer (chars / sentence avg)	102.9	115.2
Answer (sentences avg)	4.8	4.6
Answer (sentences min)	3	2
Answer (sentences max)	6	8
Answer (cites avg)	6.9	–
Answer (cites / sentence avg)	1.5	–

Table 2: Parameter comparison for ArchEHR-QA.

4. Experimental Setup

We test our approach on the proprietary GPT-5.2 model $_{\text{2025-12-11}}$[3] from OpenAI while including LLaMA-3-70B as a comparison. Our primary focus is on GPT-5.2, and we do not perform an in-depth analysis of LLaMA-3-70B. We set the temperature to 0.1 for both models. The prompts that follows the methodology outlined in Section 2 are publicly available in repository [4].

To assess the impact of clinical reasoning principles, we experiment with prompt variations: excluding reasoning principles from the prompt (†), and including reasoning principles (‡).

Evaluation metrics: The ArchEHR-QA 2026 organizers chose $F1_{micro}$ and $F1_{macro}$ as evaluation metrics. We found no publicly available evaluation script. For the results on D_{dev}, we use manually implemented calculation of $F1_{micro}$ and $F1_{macro}$. To obtain the related results, we first calculate the related $F1$ values at the sentence level and then aggregate them to the case level. The results for D_{test} split are obtained from the official evaluation script.

[3] https://openai.com/index/introducing-gpt-5-2/

[4] https://github.com/nicolay-r/ArchEHR-QA-2026-Task-4-MedEvi-NS/blob/master/src/utils_prompt.py

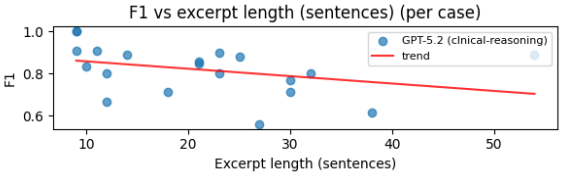

Figure 5: Detailed overview $F1_{overall}$ results for GPT-5.2 (‡) on D_{dev} based on the length of the excerpt notes in sentences

5. Result Analysis and Discussion

Table 1 presents results on ArchEHR-QA. We include results of the open-sourced model LLaMA-3-70B-instruct[5] for baseline purposes. We found that LLaMA-3-70B underperformed compared to GPT-5.2. Furthermore, we found no improvements in adopting clinical reasoning principles for LLaMA-3-70B (Uzun Bektaş et al., 2025).

Towards results on D_{dev}, adopting clinical reasoning principles (‡) for GPT-5.2 results in +2.0% improvement by $F1_{overall}$ on D_{dev}, leveraged by increased precision. Therefore, for our final submission on D_{test} we adopted GPT-5.2 ‡.

Towards results on D_{test} split, we refer to the results of team "`boris123`" as the baseline. Using GPT-5.2 ‡ improves the baseline +15.8% ($F1_{micro}$). Notably, with almost similar P_{micro} results, our system finds more citations.

According to findings from analysis of ArchEHR-QA content dataset in Section 5, *clinical excerpts* represent a significant portion once imputed in Input Description (Figure 3). To find whether the length of excerpts affects the performance, we provide case-level evaluation for our results on D_{dev} split in Figure 5. According to the related results, we found that the increment of excerpt from 10 sentences to 50 causes $\approx 10\%$ degradation of $F1_{overall}$ performance for GPT-5.2.

[5] https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

6. Conclusion

This paper proposes a clinical-reasoning principle, utilised in a form of prompt-based engineering methodology for ArchEHR-QA 2026 Subtask 4. For the given clinical interpretation of the patient question (Q_c), note excerpt (E), and answer text (A), this subtask aims to cite supporting sentences from E to sentences from A. We propose a prompt-based engineering methodology that exploits clinical reasoning principles to leverage LLM capabilities for zero-shot learning on both a proprietary model (GPT-5.2) and an open-sourced model (LLaMA-3-70B). According to our experiment on the officially provided ArchEHR-QA dataset, we found that applying clinical-reasoning principles improves $F1_{overall}$ for GPT-5.2 by +2.0%. We see those findings as a promising step for future works on the formation of advanced systems that utilise the Chain-of-Thought (CoT) paradigm.

7. Bibliographical References

Akanksha Dhamija and Arun Sharma. 2025. Prompt-enhanced question answering for clinical texts using instruction-tuned language models. In *2025 1st IEEE Uttar Pradesh Section Women in Engineering International Conference on Electrical Electronics and Computer Engineering (UPWIECON)*, pages 308–312. IEEE.

Shashi Kant Gupta, Aditya Basu, Mauro Nievas, Jerrin Thomas, Nathan Wolfrath, Adhitya Ramamurthi, Bradley Taylor, Anai N Kothari, Regina Schwind, Therica M Miller, et al. 2024. Prism: Patient records interpretation for semantic clinical trial matching using large language models. *arXiv preprint arXiv:2404.15549*.

Brian Hazlehurst, H Robert Frost, Dean F Sittig, and Victor J Stevens. 2005. Mediclass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. *Journal of the American Medical Informatics Association*, 12(5):517–529.

K Karthik, Sowmya Kamath, R Supreetha, and Ashish Katlam. 2025. Content-based medical retrieval systems with evidence-based diagnosis for enhanced clinical decision support. *Expert Systems with Applications*, 272:126678.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Chanwoo Park, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, et al. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Chenqian Le, Ziheng Gong, Chihang Wang, Haowei Ni, Panfeng Li, and Xupeng Chen. 2025. Instruction tuning and cot prompting for contextual medical qa with llms. In *2025 International Conference on Artificial Intelligence, Human-Computer Interaction and Natural Language Processing (ICAHN)*, pages 43–46. IEEE.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Jialin Liu, Fang Liu, Changyu Wang, and Siru Liu. 2025. Prompt engineering in clinical practice: tutorial for clinicians. *Journal of Medical Internet Research*, 27:e72644.

Kathleen N Lohr, Kristen Eleazer, and Josephine Mauskopf. 1998. Health policy issues and applications for evidence-based medicine and clinical practice guidelines. *Health policy*, 46(1):1–19.

Pasi Shailendra, Rudra Chandra Ghosh, Rajdeep Kumar, and Nitin Sharma. 2024. Survey of large language models for answering questions across various fields. In *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 520–527. IEEE.

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

Sarvesh Soni and Dina Demner-Fushman. 2026b. 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.

Aslıhan Uzun Bektaş, Balahan Bora, and Erbil Ünsal. 2025. Comparative evaluation of chatgpt and llama for reliability, quality, and accuracy in familial mediterranean fever. *European Journal of Pediatrics*, 184(8):491.

Laura Zakowski, Christine Seibert, and Wisconsin Selma VanEyck. 2004. Evidence-based medicine: answering questions of diagnosis. *Clinical medicine & research*, 2(1):63–69.

Xinping Zhao, Dongfang Li, Yan Zhong, Boren Hu, Yibin Chen, Baotian Hu, and Min Zhang. 2024. Seer: Self-aligned evidence extraction for retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3027–3041.