Razreshili at ArchEHR-QA 2026: Evidence Alignment via LLM Prompting and Cross-Encoder Fine-tuning

Arina Zemchyk
arina.zemchik@gmail.com

Abstract

We describe our system for Subtask 4 (Evidence Alignment) of the ArchEHR-QA 2026 shared task, which requires aligning each sentence of a clinician-authored answer to the supporting sentence(s) in a clinical note excerpt derived from MIMIC(Johnson et al., 2016). The task is challenging due to many-to-many alignment structure, answer sentences with no note support, and the semantic gap between clinical note language and answer paraphrases. We explore two approaches: few-shot chain-of-thought prompting with Qwen2.5-7B-Instruct and LoRA fine-tuning of a cross-encoder with combined InfoNCE and BCE loss. Our best system achieves a micro F1 of 67.93 on the test set.

Keywords: clinical NLP, evidence alignment, cross-encoder, parameter-efficient fine-tuning

1. Introduction

Grounded question answering over electronic health records requires not only generating accurate answers but also explicitly identifying where in the record each claim originates. The ArchEHR-QA 2026 shared task (Soni and Demner-Fushman, 2026b) formalises this challenge in four complementary subtasks. Subtask 4, which we address, requires aligning each sentence of a reference clinician-authored answer to the specific supporting sentence(s) in a clinical note excerpt, given the patient question, clinician-interpreted question, note excerpt, and reference answer as input.

The task presents several distinct challenges. First, alignment is many-to-many: a single note sentence may support multiple answer sentences, and a single answer sentence may draw on multiple note sentences. Second, answer sentences are paraphrases or aggregations of note content rather than direct extractions, requiring semantic rather than lexical matching. Third, approximately 19% of answer sentences in the development set are drawn from clinician domain knowledge rather than the note and should receive empty citation sets (Soni and Demner-Fushman, 2026a). Correctly identifying these cases is critical for precision.

We explore two complementary approaches. The first fine-tunes a cross-encoder using LoRA with a combined contrastive and calibration loss, leveraging human-annotated relevance labels from the dataset as a source of hard negatives. The second uses few-shot prompting of a quantized large language model with chain-of-thought examples that explicitly model the grounded versus inference-based distinction. Our experiments show that the LLM approach generalises better to unseen test cases, achieving micro F1 of 67.93 versus 64.44 for the fine-tuned cross-encoder.

2. Related Work

Evidence alignment is closely related to sentence-level evidence retrieval in extractive question answering and fact verification, where systems must identify minimal supporting spans for a given claim (Thorne et al., 2018; Rajpurkar et al., 2016). More recently, claim-level grounding has emerged as a distinct problem: frameworks such as eTracer (Chu et al., 2026) and AGREE (Ye et al., 2024) align individual response claims to supporting source sentences, treating grounding as a post-hoc verification step rather than retrieval. The ArchEHR-QA task differs in that alignment targets answer sentences rather than atomic claims, and some answer sentences have no note support - a distinction not present in standard extractive QA or citation generation benchmarks.

Cross-encoders have been widely used for passage relevance ranking due to their ability to jointly encode query–document pairs (Nogueira and Cho, 2019). Contrastive objectives such as InfoNCE (van den Oord et al., 2018) improve ranking quality but produce scores that are invariant to query-level shifts, making them poorly suited for threshold-based retrieval. Sheikholeslami et al. (2025) propose the Mann–Whitney (MW) loss, which directly maximises AUC by minimising binary cross-entropy over pairwise score differences, producing globally calibrated scores. Our combined InfoNCE and BCE loss addresses the same score compression problem through a simpler regularisation term, though MW loss represents a more principled future direction.

Parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022) and IA^3 (Liu et al., 2022) adapt pretrained models with few additional parameters, making them attractive for shared tasks with limited training data.

Large language models have demonstrated strong performance on clinical reasoning tasks

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 524–529

via few-shot prompting (Singhal et al., 2023), and chain-of-thought prompting improves multi-step reasoning (Wei et al., 2022). In the 2025 iteration of this shared task, the runner-up system (Bogireddy et al., 2025) used DSPy-optimised prompts with self-consistency voting for sentence-level evidence identification, outperforming fine-tuned retrieval approaches by a substantial margin - consistent with our finding that LLM prompting generalises better than cross-encoder fine-tuning in low-data clinical settings. Our prior work (Zemchyk, 2025) explored contrastive fine-tuning of a bi-encoder for retrieval-augmented answer generation in the same task family; here we extend this to evidence alignment with a cross-encoder and a combined ranking and calibration loss.

3. System Description

3.1. Data and Task Setup

The development set contains 20 cases with 95 answer sentences, of which 77 have gold citations and 18 are inference-based with no supporting note sentences. Note sentences are annotated with relevance labels: *essential*, *supplementary*, and *not-relevant*. The test set contains 147 cases. Evaluation uses micro F1 as the primary metric.

3.2. Approach 1: Zero-shot Cross-encoder Baseline

We evaluated four candidate cross-encoders on the development set: `ms-marco-MiniLM-L-6-v2` (general, fast), `ms-marco-MiniLM-L-12-v2` (general, larger), `pritamdeka/S-PubMedBert-MS-MARCO` (biomedical domain-adapted), and `ncbi/MedCPT-Cross-Encoder` (biomedical). Despite the biomedical models being domain-matched to the clinical task, `ms-marco-MiniLM-L-12-v2` achieved the highest development macro F1 after threshold tuning, consistent with prior findings that MS MARCO cross-encoders transfer strongly to out-of-domain retrieval (Thakur et al., 2021; Rosa et al., 2022). We therefore selected it as the base model for both the zero-shot baseline and subsequent fine-tuning.

For inference, we applied a two-stage strategy: if the maximum sigmoid score across all note sentences falls below a no-citation threshold t_{nc}, the answer sentence is assigned an empty citation set; otherwise, all note sentences scoring above a citation threshold t_c are returned, with a fallback to the top-scoring sentence. Thresholds were tuned on the development set ($t_{nc} = 0.2$, $t_c = 0.4$), achieving a development macro F1 of 0.723. Two implementation details were critical: applying sigmoid to raw logits before thresholding, and correctly scoring empty-vs-empty citation pairs as F1 = 1.0 rather than 0.0.

3.3. Approach 2: Cross-Encoder Fine-tuning with Two-Stage Inference

We fine-tuned `ms-marco-MiniLM-L-12-v2`(Nguyen et al., 2016) using LoRA (r=4, α=8) with a two-stage inference pipeline that separates the citation decision from evidence ranking, allowing each to be calibrated independently.

Stage 1: Citation Decision. For each answer sentence, we compute sigmoid scores across all note sentences. If the maximum score falls below a no-citation threshold t_{nc}, the sentence is predicted as inference-based and assigned an empty citation set.

Stage 2: Evidence Ranking. For sentences passing Stage 1, all note sentences scoring at or above a citation threshold t_c are cited, with a fallback to the single highest-scoring sentence. This stage is where fine-tuning contributes most directly - the LoRA-adapted model is trained to rank gold citation sentences above hard negatives.

Training Data. We constructed 138 contrastive groups from all 20 development cases. For each (answer sentence, gold citation) pair, hard negatives were sampled using a three-tier priority: (1) *supplementary* sentences, human-annotated as relevant but not essential - the most confusable negatives available; (2) sentences within a $\pm$2 position window of the gold citation; (3) other *not-relevant* sentences. Easy negatives (section headers, sentences under 6 words) were excluded. Inference-based answer sentences were excluded from training entirely, as including them consistently degraded the citation decision in Stage 1.

Loss Function. We trained with a combined loss:

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \lambda \cdot \mathcal{L}_{\text{BCE}} \quad (1)$$

where $\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(s_+/\tau)}{\sum_i \exp(s_i/\tau)}$ with temperature $\tau = 0.2$, and $\mathcal{L}_{\text{BCE}}$ is binary cross-entropy with the positive labelled 1.0 and negatives 0.0, weighted by $\lambda = 0.3$. InfoNCE trains Stage 2 ranking, while BCE anchors absolute score magnitudes to prevent compression that would destabilise the Stage 1 threshold.

Models were trained for 8 epochs with a learning rate of 2×10^{-4}, linear warmup over 10% of steps, and AdamW optimiser with weight decay 0.01.

Threshold Tuning. Both thresholds were tuned jointly on the development set via grid search over $t_{nc} \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$ and $t_c \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, optimising macro F1. The fine-tuned model required higher thresholds ($t_{nc} = 0.6$, $t_c = 0.7$) than the zero-shot model ($t_{nc} = 0.2$, $t_c = 0.4$), reflecting the upward shift in score magnitudes produced by the BCE calibration term.

An initial attempt used InfoNCE alone without sigmoid calibration, treating thresholds as operating on raw logits; this achieved micro F1 of 59.95 with high precision (77.16) but poor recall (49.02), confirming that uncalibrated scores make threshold tuning unreliable.

All experiments were run on a single T4 GPU.

3.4. Approach 3: LLM Few-shot Prompting

We used `Qwen2.5-7B-Instruct`(Qwen et al., 2025) running in 4-bit quantization via bitsandbytes on a single T4 GPU. Inference over the full test set of 147 cases required approximately 30 minutes, compared to under one minute for the fine-tuned cross-encoder including threshold tuning on the development set. The full prompt template is provided in Appendix A.

The prompt included three few-shot examples with chain-of-thought reasoning:

- a single-citation case
- a multi-citation case
- an inference-based case with empty citations

The third example was critical; without it the model consistently over-cited inference sentences. The prompt explicitly distinguished sentences that *directly state* the answer from sentences that are merely related.

Outputs were constrained to JSON format.

4. Results

Table 1 presents the performance of the evaluated systems on the ArchEHR-QA Subtask 4 test set. The Qwen-based system outperformed all cross-encoder variants. While the LoRA fine-tuned model substantially improved recall compared to the early cross-encoder baseline (65.66 vs. 49.02), this came at the cost of reduced precision. The Qwen approach achieved the best overall balance between precision and recall, resulting in the highest micro F1 score.

5. Discussion

5.1. LLM Prompting vs. Cross-Encoder Fine-tuning

Despite being fine-tuned on the task with human-annotated hard negatives, the LoRA cross-encoder (micro F1 64.44) underperformed the few-shot LLM (micro F1 67.93). We attribute this to two complementary factors: the nature of the alignment task itself, and the limitations imposed by the small development set.

Evidence alignment requires determining whether an answer sentence is *semantically entailed by* a note sentence - a reasoning operation that goes beyond surface similarity. Cross-encoders pre-trained on MS MARCO learn to match queries to relevant passages, but clinical entailment involves paraphrase, aggregation, and implicit inference that differ structurally from passage retrieval. Qwen2.5-7B, by contrast, can reason explicitly about whether a claim is *directly stated* in the note or represents clinician inference, as demonstrated by its ability to correctly handle inference-based sentences when provided with chain-of-thought examples.

From a practical standpoint, the cross-encoder approach offers a substantial efficiency advantage: the full pipeline including development set threshold tuning runs in under one minute on a T4 GPU, compared to approximately 30 minutes for LLM inference over the same number of cases. This motivates further work on closing the performance gap through better contrastive training objectives and calibration methods.

5.2. Representation Collapse Under Fine-tuning

A recurring failure mode across our fine-tuning experiments was score compression: after training, sigmoid scores for all note sentences clustered in a narrow range, shifting the optimal citation threshold from 0.4 (zero-shot) toward 0.3 or lower. In severe cases, the model collapsed entirely, predicting empty citations for all answer sentences (NoCiteAcc = 1.0, F1 $\approx$ 0.35).[1]

We hypothesise that this reflects a form of representation collapse specific to small contrastive datasets. With only 138 training pairs across 20 cases, the model encounters the same note sentences repeatedly across different training examples. The LoRA layers adapt to reduce loss on these specific sentence pairs, but in doing so compress the representation space: note sentences

[1] NoCiteAcc is the proportion of inference-based answer sentences (those with empty gold citation sets) correctly predicted as having no supporting note sentence.

System	Precision	Recall	F1
Early cross-encoder (no sigmoid fix)	**77.16**	49.02	59.95
LoRA + combined loss	63.27	**65.66**	64.44
Qwen2.5-7B few-shot (best)	73.78	62.95	**67.93**

Table 1: Test set performance on ArchEHR-QA Subtask 4.

that were distinguishable under the pre-trained model become more similar in the fine-tuned embedding space, making threshold-based discrimination harder rather than easier. This is consistent with known collapse phenomena in contrastive learning (Jing et al., 2022), where insufficient diversity in training pairs leads to degenerate representations.

Adding a BCE calibration term to the InfoNCE loss partially mitigated this by anchoring absolute score magnitudes, improving dev macro F1 from 0.704 to 0.760. However, the large gap between dev macro F1 (0.760) and test micro F1 (64.44) suggests the calibration was itself overfit to the score distribution of the 20 dev cases rather than generalising to the 147 test cases.

5.3. Threshold Tuning as the Dominant Factor

Our two-stage inference design separates the citation decision (Stage 1) from evidence ranking (Stage 2), with fine-tuning primarily targeting the latter. However, our results suggest that threshold calibration for Stage 1 contributed more to overall performance than the ranking improvements from fine-tuning.

Cross-encoder with optimised thresholds ($t_{nc} = 0.2$, $t_c = 0.4$) achieved dev macro F1 of 0.723, comparable to or exceeding several fine-tuned variants. The first submitted model, which lacked sigmoid calibration and used untuned thresholds, scored only 59.95 micro F1 - well below what the zero-shot model with proper calibration would have achieved. Applying sigmoid transformation to raw logits and separately tuning t_{nc} and t_c on the development set was the single largest source of improvement across all experiments.

Fine-tuning improved Stage 2 ranking as intended(recall increased substantially from the first to final submission (49.02 $\rightarrow$ 65.66)) but at the cost of precision (77.16 $\rightarrow$ 63.27). The model learned to cite more aggressively in Stage 2 without learning when to abstain in Stage 1, suggesting that the ranking and abstention objectives interfere when trained jointly on a small dataset. The BCE term partially decoupled these by calibrating absolute scores, but the Stage 1 threshold still had to shift substantially ($t_{nc} : 0.2 \rightarrow 0.6$) to compensate for the upward score drift introduced by training.

5.4. The Inference Sentence Problem

Correctly identifying inference-based answer sentences - those with empty citation sets - proved to be a surprisingly tractable problem for the zero-shot cross-encoder. At $t_{nc} = 0.2$, the model correctly identified 15 of 18 such sentences on the dev set (NoCiteAcc = 0.833) without any task-specific training, simply because these sentences have low semantic overlap with any note content. The three failures involved sentences that, while inference-based, happened to be topically related to note content (e.g., discussing a procedure mentioned in the note). Fine-tuning degraded this capability in most experiments, likely because excluding inference sentences from training left the model without signal about *when* to abstain from citing.

6. Conclusion

We presented two approaches for clinical evidence alignment: cross-encoder fine-tuning with a combined ranking and calibration loss, and LLM few-shot prompting.

The LLM approach achieved the best performance (micro F1 67.93), outperforming fine-tuned cross-encoders on a task that requires semantic reasoning over clinical text. However, the cross-encoder approach is approximately 30$\times$ faster at inference, making it more suitable for deployment settings where low latency matters. Future work could explore hybrid approaches that retain the reasoning quality of LLMs while reducing inference cost, such as using LLM predictions as soft labels for cross-encoder distillation, or using cross-encoder scores as lightweight features within the LLM prompt.

7. Limitations

Results are based on a small development set of 20 cases, which limits the reliability of threshold tuning and ablation conclusions. Additionally, the Qwen model was run in 4-bit quantization, which may reduce performance compared to full precision.

8. Bibliographical References

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. Association for Computational Linguistics.

Bohao Chu, Qianli Wang, Hendrik Damm, Hui Wang, Ula Muhabbek, Elisabeth Livingstone, Christoph M. Friedrich, and Norbert Fuhr. 2026. etracer: Towards traceable text generation via claim-level grounding. *arXiv preprint arXiv:2601.03669*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. 2022. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*.

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

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated MAchine reading COmprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*, volume 1773. CEUR-WS.org.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Guilherme Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. In defense of cross-encoders for zero-shot retrieval. *arXiv preprint arXiv:2212.06121*.

Nima Sheikholeslami et al. 2025. Optimizing what matters: AUC-driven learning for robust neural retrieval. *arXiv preprint arXiv:2510.00137*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

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.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, pages 809–819. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Xi Ye, Ruoxi Sun, Sercan Ö. Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6237–6251. Association for Computational Linguistics.

Arina Zemchyk. 2025. razreshili at ArchEHR-QA 2025: Contrastive fine-tuning for retrieval-augmented biomedical QA. In *Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP) at ACL 2025*.

A. LLM Prompt Template

We design a structured prompt consisting of a system instruction and a user prompt with few-shot examples to guide the model in identifying evidence sentences.

System Prompt

```
You are a clinical NLP expert. Your task is to identify which sentences from a clinical note directly support a given answer sentence.

You must respond with ONLY a JSON object - no explanation, no markdown, no extra text.

Follow these rules strictly:
- INCLUDE a sentence if it directly states the same information as the answer sentence (verbatim or paraphrased). - INCLUDE all sentences that each contribute a distinct piece of information present in the answer sentence. - DO NOT include a sentence just because it is adjacent to or on the same topic as a supporting sentence - it must independently state part of the answer. - DO NOT include a sentence that only provides context, background, or consequence of what the answer states. - Return [] if no note sentence directly states the answer content.
```

User Prompt with Few-Shot Examples

```
-- [EXAMPLE 1 (single supporting sentence; enforces precision, avoids over-citation)]
--
-- [EXAMPLE 2 (multi-sentence candidate set; enforces selecting only directly supporting evidence, ignoring adjacent/context sentences)]
--
-- [EXAMPLE 3 (no supporting evidence; enforces empty output when answer is not grounded in the note)] --
-- END EXAMPLES --

Now do the same for:

Answer sentence:
"<ANSWER_SENTENCE>"

Clinical note sentences:
<NOTE_BLOCK>

Which note sentence IDs directly support the answer sentence?

Respond with ONLY this JSON:
"evidence_id": ["<id>", ...]
```