razreshili at ArchEHR-QA 2025: Contrastive Fine-Tuning for Retrieval-Augmented Biomedical QA

Arina Zemchyk
arina.zemchik@gmail.com

Abstract

We present a retrieval-augmented system for the ArchEHR-QA 2025 shared task, which focuses on generating concise, medically accurate answers to clinical questions based on a patient's electronic health record (EHR). A key challenge is following a strict citation format that references relevant sentence IDs. To improve retrieval, we fine-tuned an all-MiniLM-L6-v2 embedding model using contrastive learning on over 2,300 question–sentence triplets, with DoRA for efficient adaptation. Sentences were selected using cosine similarity thresholds and passed into a quantized Mistral-7B-Instruct model along with a structured prompt. Our system achieved similar relevance to the baseline but lower overall performance (19.3 vs. 30.7), due to issues with citation formatting and generation quality. We discuss limitations such as threshold tuning, prompt-following ability, and model size, and suggest future directions for improving structured biomedical QA.

1 Introduction

The ArchEHR-QA 2025 shared task focuses on answering medical questions based on a patient's electronic health record (EHR) (Soni and Demner-Fushman, 2025b). Each answer must be short, medically accurate, and include in-text citations using sentence IDs from the patient history (e.g., |1,2|). This makes the task challenging, especially due to the length and complexity of clinical records and the strict output formatting rules.

Our approach follows a retrieval-augmented pipeline. First, we fine-tune an embedding model to better identify relevant sentences in the patient's history. Then, we pass the selected sentences, together with the question, into a generative model (Mistral-7B[1]) that produces the answer.

Although our method did not outperform the baseline, it achieved comparable relevance score. Most of the performance gap came from formatting issues and citation errors in the generated text, which we analyze in this paper. We also discuss the challenges of tuning models with limited data and propose directions for improvement.

2 Methodology

In order to improve the accuracy and relevance of cited sentences in generated answers, the main focus of the proposed system is a domain-adapted embedding model, which can capture the nuances of a biomedical domain.

2.1 Overview

The approach consists of three main steps: (1) fine-tuning an embedding model on the development set of the shared task dataset (Soni and Demner-Fushman, 2025a), (2) selecting relevant and supplementary context sentences based on cosine similarity thresholds, and (3) generating answers using a quantized generative model (Mistral-7B) with in-context citations.

2.2 Embedding Model Fine-Tuning

To accurately retrieve relevant sentences from the patient's history, we fine-tuned the `all-MiniLM-L6-v2`[2] model with contrastive objective using DoRA (Mao et al., 2024), a parameter-efficient fine-tuning method that extends LoRA (Hu et al., 2022). DoRA improves learning capacity and training stability of LoRA, making it particularly suitable in settings with limited training data and computational resources. Additionally, parameter-efficient tuning mitigates the issue of catastrophic forgetting (Goodfellow et al., 2013), where the pretrained model loses its

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

BioNLP 2025 Shared Tasks, pages 160–164

original knowledge during full fine-tuning. Details about DoRA setup can be found in Appendix A.

We constructed a dataset of 2,582 triplets from the development set (Soni and Demner-Fushman, 2025a), where each triplet consisted of:

- **Anchor:** a clinical question,
- **Positive:** a sentence labeled as "essential" or "supplementary" to a given clinical question,
- **Negative:** a sentence labeled as "not relevant" to a given clinical question.

The dataset was split into 2,341 training and 241 validation triplets to monitor performance.

We used the `Trainer`[3] from the `SentenceTransformer` library with `MultipleNegativesRankingLoss` (analogous to InfoNCE loss (Oord et al., 2018)) as the training objective. In this setup, negatives were treated as in-batch negatives, with the explicit negative in each triplet acting as a hard negative. Training was run for 50 epochs with the following key hyperparameters: batch size of 64 (train) and 128 (eval), learning rate of 1e-4, warmup ratio of 0.1, and no-duplicates batch sampling (beneficial for in-batch negative mining).

To monitor training, we evaluated embedding quality using alignment and uniformity metrics (Figure 1). Alignment is a metric that measures the closeness of positive pairs representations. Uniformity, on the other hand, depicts how well the embeddings are ditstributed on a unit hypersphere. These metrics were introduced by Wang and Isola (2020) and provide insights into how well the fine-tuned model clustered relevant sentences closer to their corresponding questions while maintaining separation from irrelevant ones.

2.3 Threshold Selection for Relevance

To define a threshold for sentence relevance, we embedded both the clinical questions and patient history sentences using the fine-tuned model and computed cosine similarity scores. Thresholds were empirically determined by testing similarity values between 0.0 and 1.0 (in increments of 0.01) on the development set, selecting the threshold that produced the highest F1 score for identifying "relevant" sentences:

- **Relevant:** cosine similarity ≥ 0.25
- **Supplementary:** $0.20 \leq$ cosine similarity < 0.25
- **Irrelevant:** cosine similarity ≤ 0.20

During answer generation on the test set, if no sentences met the "relevant" or "supplementary" criteria (i.e., all sentences were classified as "irrelevant"), the full patient history was used as context.

2.4 Generative QA Module

For answer generation, we used a quantized `Mistral-7B-Instruct-v0.2`[4] model, selected due to computational constraints. The prompt was structured into three segments:

1. **Instruction Header:** a detailed instruction block framing the task, e.g., "You are a medical assistant tasked with answering patient questions using provided case information. After each factual claim, cite supporting sentences in the format |id| or |id1, id2|. Limit the answer to 75 words."
2. **Context:** a concatenation of the retrieved relevant and supplementary sentences, each labeled with its sentence ID for proper referencing.
3. **Clinical Question:** the specific question to be answered.

The prompt also included an explicit example demonstrating correct citation style and answer formatting, to help enforce the desired output pattern. Despite these explicit instructions, we observed that the generative model frequently struggled to fully comply with strict citation formatting and word count limits, highlighting typical challenges in controlling large language models.

The full prompt template used in this work is provided in Appendix C.

2.5 Reflections

The final system's underperformance relative to the baseline may be from two main factors: (1) intrinsic weaknesses of the generative model in structured QA and (2) potentially over-restrictive relevance thresholds, which may have omitted valuable context. The small development set size also limited threshold generalizability.

[3] https://sbert.net/docs/package_reference/sentence_transformer/trainer.html

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

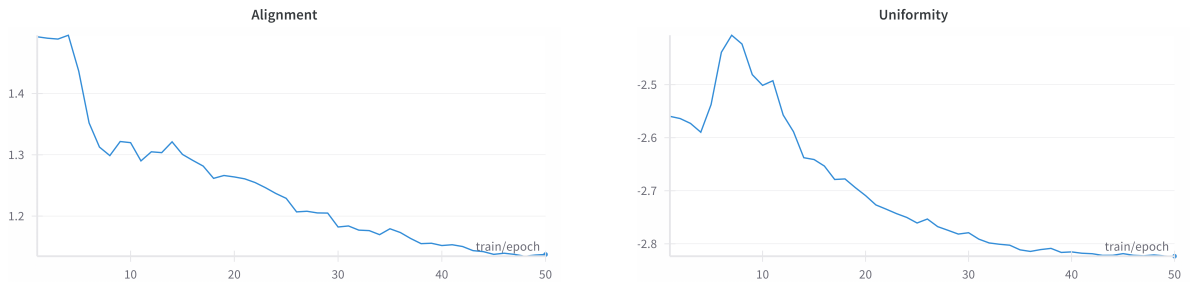

Figure 1: Training progression visualized through alignment (left) and uniformity (right) metrics on a validation set.

3 Results

We submitted one main official run for the ArchEHR-QA task using the `Mistral-7B-Instruct-v0.2`-based system. Furthermore, we experimented with another generative model `meta-lama-3-8B-Instruct`[5], but its performance was slightly lower (overall score: 19.2) and therefore it is not considered in this section. Table 1 reports the scores of our main run compared to the organizers' baseline.

Metric	Baseline	Ours
Overall	30.7	19.3
Overall Factuality	33.6	13.5
Overall Relevance	27.8	25.2
Strict Precision (micro)	71.6	36.8
Strict Recall (micro)	21.9	8.2
Strict F1 (micro)	33.6	13.5
Lenient Precision (micro)	77.0	39.7
Lenient Recall (micro)	22.3	8.4
Lenient F1 (micro)	34.6	13.9
Strict Precision (macro)	77.4	49.6
Strict Recall (macro)	31.5	14.5
Strict F1 (macro)	39.0	19.0
Lenient Precision (macro)	83.0	53.8
Lenient Recall (macro)	30.8	13.6
Lenient F1 (macro)	39.9	19.1
BLEU	0.1	0.4
ROUGE-Lsum	15.2	16.8
SARI	47.8	45.8
BERTScore	20.5	19.9
AlignScore	57.7	43.9
MEDCON (UMLS)	25.6	24.5

Table 1: Performance comparison between the baseline (organizers) and our system (razreshili) on the ArchEHR-QA test set.

Our best submission did not outperform the baseline in most official metrics, but achieved a comparable relevance score (25.2 vs. 27.8 overall relevance) and slightly higher ROUGE-Lsum (16.8 vs. 15.2) and BLEU (0.4 vs. 0.1).

3.1 Error Analysis

We conducted a detailed error analysis with the following findings:

- **Citation format errors:** Despite explicit prompt engineering, some generated answers failed to follow the required citation format (`|sent_id|`, e.g., `|1|`). This often happened when the context was complex or included many sentences. Typical mistakes included separating citations incorrectly (e.g., `|1|, |12|` instead of `|1,12|`), breaking them across lines (e.g., `|3|\n\n|2,6|` instead of `|3,2,6|`), or separating with a dot (e.g., `|6|. |1,3|` instead of `|6,1,3|`). These formatting issues might have contributed to lower scores in strict citation metrics.
- **Word limit violations:** Of 100 cases, 14 generated responses exceeded the 75-word limit. We observed that these violations were more common in truncated cases, where the context length was substantially longer: truncated cases had on average 7.4 relevant and 5.3 supplementary sentences, compared to 3.6 and 2.0 in non-truncated cases. This suggests that longer, information-rich contexts increased the likelihood of the model producing overlength answers.

4 Discussion

Our method did not outperform the baseline, but it helps show where smaller generative models struggle in biomedical question answering.

Even though adding relevant and supplementary sentences to the prompt helped us reach a similar

[5] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

relevance score, other scores like citation accuracy and factual correctness were much lower. This means that better sentence retrieval alone is not enough—the model also needs to follow strict rules for format and content.

Smaller models like Mistral-7B and Meta-Llama-3-8B often failed to follow the required citation format or stay under the 75-word limit. In contrast, larger models like LLaMA 3.3 70B, which were used in the baseline system, are better at following instructions and producing more accurate answers. While we used a retrieval-augmented setup to shorten the context and focus the model on relevant sentences, newer models like LLaMA-3-8B or Mistral-7B support longer inputs and could process the full patient history directly. We didn't try this due to limited resources, but it could be a strong and simpler baseline for future work.

In future work, combining better retrieval with larger or more fine-tuned generative models may help improve performance on this type of task.

5 Limitations

Our approach has several limitations:

- **Small dev set:** The development set was small, which made it hard to properly adapt a sentence embedding model to a complex medical domain.
- **Strict thresholds:** The fixed similarity thresholds for selecting relevant and supplementary sentences may have removed useful context, especially for more difficult questions.
- **Generative model constraints:** We used a quantized version of Mistral-7B due to hardware limitations. While fast and memory-efficient, this model often failed to follow citation and length constraints, limiting the effectiveness of our retrieval pipeline.
- **No fine-tuning of the generator:** The generator was used as-is with prompt instructions. We didn't fine-tune it on this task, which likely hurt citation accuracy.
- **Prompt sensitivity:** Despite careful prompt design, the model often ignored citation formatting rules. This suggests that prompt-only control may be insufficient for tasks with strict output requirements.
- **No baseline for smaller embedding model:** We did not compare our fine-tuned embedding model against the original (non-adapted) version. This limits our ability to directly measure the contribution of contrastive fine-tuning to retrieval performance.

References

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

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*.

Yulong Mao, Kaiyu Huang, Changhao Guan, Ganglin Bao, Fengran Mo, and Jinan Xu. 2024. DoRA: Enhancing parameter-efficient fine-tuning with dynamic rank distribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11662–11675, Bangkok, Thailand. 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*.

Sarvesh Soni and Dina Demner-Fushman. 2025a. A dataset for addressing patient's information needs related to clinical course of hospitalization. *arXiv preprint*.

Sarvesh Soni and Dina Demner-Fushman. 2025b. Overview of the archehr-qa 2025 shared task on grounded question answering from electronic health records. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, pages 9929–9939. PMLR.

A DoRA Fine-Tuning Configuration

```
config = LoraConfig(
    target_modules = ["value", "query"],
    use_dora=True,
    r=16,
    lora_alpha=32,
    lora_dropout=0.01,
```

```
    bias="none",
)
```

B Generation Parameters

The following parameters were used during answer generation with Mistral-7B-Instruct-v0.2:

```
generation_kwargs = {
   "pad_token_id": tokenizer.eos_token_id,
    "max_new_tokens": 512,
    "temperature": 0.2,
    "top_p": 0.95,
    "do_sample": True,
}
```

C Prompt Template

```
You are a medical assistant tasked with answering patient questions using provided
case information.
Rules:
- After every factual claim, cite the supporting sentence(s) in the format |id| or
|id1, id2|.
- Group citations if multiple sentences support the same claim (e.g., |1,2|).
- Do not create a 'References' section.
- Limit the answer to 75 words or fewer.
- Only use the provided sentences; do not hallucinate facts.
- Write clearly, medically accurately, and concisely.

Example:
Evidence:
1. The patient has alcoholic cirrhosis.
2. He has advanced hepatic encephalopathy.
3. His renal function is deteriorating.

Question:
What is the patient's prognosis?

Answer:
The patient's prognosis is poor due to alcoholic cirrhosis |1|, advanced hepatic
encephalopathy |2|, and worsening renal function |3|.

Patient Question:
{QUESTION}

Relevant Information:
- {Sentence text} |{sentence_id}|
...

Supplementary Information (less directly relevant but possibly helpful):
- {Sentence text} |{sentence_id}|

Now, based on the evidence, write your answer:
```