UniBuc-SB at ArchEHR-QA 2025: A Resource-Constrained Pipeline for Relevance Classification and Grounded Answer Synthesis

Sebastian Balmuș[1,2], Bogdan Dura[1,2], and Ana-Sabina Uban[1,3]

[1]Faculty of Mathematics and Computer Science, University of Bucharest
[2]National Institute for Research and Development in Informatics - ICI Bucharest
[3]Human Language Technologies Research Center, University of Bucharest

Abstract

We describe the UniBuc-SB submission to the ArchEHR-QA shared task, which involved generating grounded answers to patient questions based on electronic health records. Our system exceeded the performance of the provided baseline, achieving a higher performance in generating contextually relevant responses. Notably, we developed our approach under constrained computational resources, utilizing only a single NVIDIA RTX 4090 GPU. We refrained from incorporating any external datasets, relying solely on the limited training data supplied by the organizers. To address the challenges posed by the low-resource setting, we leveraged off-the-shelf pre-trained language models and fine-tuned them minimally, aiming to maximize performance while minimizing overfitting.

1 Introduction

The ArchEHR-QA shared task (Soni and Demner-Fushman, 2025b) focuses on advancing automated question answering systems that can generate grounded responses using electronic health records (EHRs). With the increasing use of patient portals, clinicians are increasingly challenged by the volume of patient inquiries. Automating the response process aims to reduce this workload by providing quick and accurate answers to patients. The task provides realistic patient queries along with clinical notes, requiring the systems to generate answers based on the EHR excerpts provided. This setting not only tests the ability to handle limited data, but also emphasizes the need for accurate medical language understanding and effective information retrieval.

Developing effective question answering (QA) systems for the medical domain presents distinct challenges, particularly when working with limited data and computational resources. The development dataset (Soni and Demner-Fushman, 2025a; Johnson et al., 2023a,b) provided was relatively small, consisting of only 20 distinct medical cases, while the test dataset consisted of 100 medical cases. This data limitation increased the risk of overfitting and restricted the potential for extensive training. Additionally, the complexity of medical language requires systems to accurately interpret nuanced terminology and context. To address these challenges, we adopted a resource-efficient approach, using a single NVIDIA RTX 4090 GPU and adhering strictly to the provided dataset, without incorporating any external data. Our system leveraged pre-trained language models to compensate for the data limitations, applying minimal fine-tuning to adapt them to the medical QA task. This strategy aimed to balance computational efficiency with performance, allowing our system to effectively generate grounded answers despite the small dataset size. Our results demonstrate that even under these constraints, our approach exceeded the baseline, highlighting the effectiveness of strategic model selection and fine-tuning in low-resource settings.

The remainder of this paper is organized as follows. Section 2 discusses related work, focusing on prior approaches to medical question answering and low-resource NLP systems. Section 3 details our system architecture, including data preprocessing, model selection, and training procedures. Section 4 presents the results of our system compared to the baseline, accompanied by a thorough analysis of its performance. Finally, Section 5 concludes the paper by summarizing our findings, highlighting limitations, and suggesting directions for future work.

2 Related Work

Recent advancements in EHR question answering (QA) systems have focused on improving information retrieval accuracy while mitigating hallucinations and enhancing interpretability. Bardhan

BioNLP 2025 Shared Tasks, pages 62–68

et al. (2023) provide a comprehensive review of EHR QA research, identifying the emrQA dataset as the primary resource and emphasizing the need for standardized evaluation metrics to facilitate consistent benchmarking.

In response to the need for robust evaluation frameworks, EHRNoteQA (Kweon et al., 2024) was introduced as a benchmark designed to assess Large Language Models (LLMs) on patient-specific questions derived from MIMIC-IV (Johnson et al., 2023a) discharge summaries. The dataset includes both open-ended and multiple-choice questions and has been used to systematically evaluate 27 LLMs, highlighting the variability in model performance across different question types.

Addressing the challenge of querying structured EHR data, quEHRy (Soni et al., 2023) employs natural language interfaces to translate clinician queries into structured database queries, facilitating more intuitive data access and emphasizing interpretability.

In the context of ensemble learning, Romero et al. (2025) demonstrate that leveraging multiple BERT-based encoders significantly improves medication-related named entity recognition (NER) across dosage, route, and strength attributes. This approach aligns with our system design, which employs multi-model architectures to capture complementary error patterns.

Finally, Sohn et al. (2024) introduce RAG2, a retrieval-augmented generation framework that prioritizes rationale-driven query formulation and evidence sampling to reduce hallucinations. Their findings underscore the value of multi-pass answer generation and rationale-centric retrieval, both of which inform our system's evidence-grounding strategy.

3 System Description

Our system is structured as a modular pipeline composed of three main components: preprocessing, relevance classification and answer generation, as shown in Figure 1. The pipeline is designed to process input data consisting of electronic health records (EHRs) and clinician question, transforming them into structured data for downstream processing. The preprocessing component structures the input data, which is then fed into the relevance classification module to identify relevant sentences. The identified sentences are subsequently processed in the answer generation module, which consists of three sequential steps: generation, grounding, and post-processing. The final output is a contextually grounded response tailored to the question.

3.1 Preprocessing

The preprocessing stage structures raw data into a format suitable for downstream tasks. Each medical case is divided into sentences labeled as essential, supplementary, or irrelevant based on their relevance to the clinician's query. The query is incorporated as contextual input for relevance classification. Each record includes a case ID, sentence ID, sentence text, query, and relevance label, ensuring consistency in data handling.

To prevent data leakage, the dataset is split at the case level, maintaining label distribution across training and testing sets. Relevance labels are then binarized, with essential and supplementary sentences labeled as 1 and irrelevant sentences as 0, simplifying the classification task.

3.2 Relevance Classification

The relevance classification component is responsible for identifying sentences within the input data that are relevant or not to the clinician's query. This step is critical in filtering out irrelevant content and ensuring that subsequent processing stages focus solely on clinically pertinent information.

To accomplish this, we employ an ensemble classifier composed of four pre-trained language models. Each model is fine-tuned for binary relevance classification, distinguishing between relevant and irrelevant content. The selected models include BERT[1] (Devlin et al., 2019), Bio_ClinicalBERT[2] (Alsentzer et al., 2019), BlueBERT[3] (Peng et al., 2019), and MedEmbed[4] (Balachandran, 2024). This combination allows us to leverage both general-domain language understanding through BERT and domain-specific medical knowledge through the clinical and biomedical models, ensuring that the classifier can effectively handle both general and specialized content within the EHR data.

[1] https://huggingface.co/google-bert/bert-base-uncased

[2] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

[3] https://huggingface.co/bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12

[4] https://huggingface.co/abhinand/MedEmbed-large-v0.1

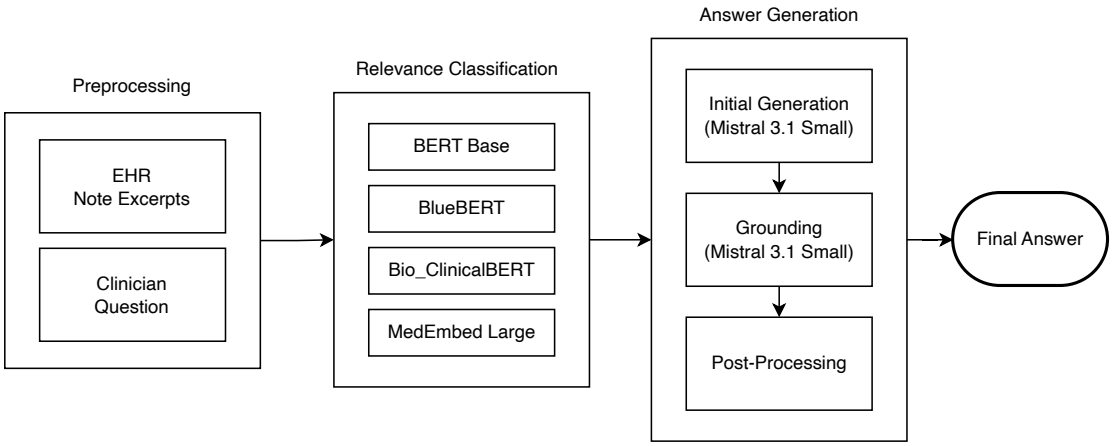

Figure 1: Our proposed system architecture, structured as a modular pipeline involving preprocessing, relevance classification, and answer generation. The entire system is designed to operate within a 24GB VRAM limit.

To address the class imbalance present in the dataset, we adopt the focal loss (Lin et al., 2020) as the objective function during training. Originally developed for dense object detection in computer vision, focal loss adjusts the contribution of each sample to the overall loss based on its classification difficulty. Specifically, it down-weights the contribution of well-classified samples and focuses more on harder-to-classify examples, mitigating the impact of the overrepresented non-relevant class in our dataset.

The training setup employs a learning rate of 5×10^{-5}, with a linear warm-up schedule comprising 10% of the total training steps. The training process is conducted over seven epochs, with a batch size of 16 samples per step. Model checkpoints are saved at each epoch, with the best model based on validation F1 score selected as the final model.

Upon completing individual model training, the predictions from each model are aggregated to form the ensemble output. For each input sentence, the relevance label is determined by majority voting, wherein the label receiving the highest number of votes across all models is selected as the final prediction. This ensemble strategy leverages the strengths of each model, reducing the impact of individual model biases and enhancing overall classification robustness.

The output of the relevance classification step serves as the input to the subsequent answer generation module. Only sentences that are classified as relevant to the clinician's query are retained.

3.3 Answer Generation

The generation stage begins by compiling the relevant sentences identified during the classification phase. Each sentence is formatted with its unique identifier and presented in a structured evidence list. This evidence list is then combined with the clinician's question to form a comprehensive input prompt for the generation model.

For response generation, we employ the Mistral Small 3.1 language model (Mistral AI, 2025), which is designed to handle large-scale language tasks with a compact yet powerful architecture. The model is loaded using the Ollama interface with the default parameters, which provides a seamless integration for inference and allows efficient model deployment without extensive modification of the original architecture. This integration facilitates the use of the model within the existing pipeline without exceeding the 24GB VRAM limit imposed by the RTX 4090 GPU, ensuring that the entire system remains computationally feasible.

The input prompt instructs the model to generate a concise response that addresses the clinician's query while adhering to a specified word limit. If the generated response exceeds the maximum word limit of 70 words, the generation step is repeated with a modified prompt that instructs the model to produce a more succinct version of the response. This iterative refinement process ensures that the output remains within acceptable length constraints without compromising informativeness.

Following response generation, the grounding step is employed to verify and reinforce the generated response by explicitly referencing relevant

evidence from the input sentences. This step mitigates the risk of unsupported claims and enhances the factual accuracy of the output, aligning it with the context provided by the EHR data.

The final post-processing step involves correcting formatting inconsistencies, such as erroneous citations or incomplete sentences. Additionally, the post-processing script ensures that the output structure is consistent across cases, aligning with the required submission format. This step is crucial for maintaining the overall quality and coherence of the generated responses.

4 Evaluation

The evaluation is conducted on a test set of 100 medical cases, focusing on factuality and relevance. Factuality is assessed using Precision, Recall, and F1 scores by comparing generated evidence citations with the ground truth.

Factuality evaluation includes Strict and Lenient modes. Strict considers only 'essential' sentence citations, while Lenient also includes 'supplementary' sentences, allowing for more flexibility.

Relevance is evaluated using BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), SARI (Xu et al., 2016), BERTScore (Zhang et al., 2020), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023), assessing both linguistic quality and clinical grounding.

The overall score is the average of the Strict Citation F1 score (Factuality) and a composite Relevance score, calculated by normalizing and averaging the individual metric scores.

4.1 Evaluation Results

Table 1 presents the evaluation results of our system across various metrics, focusing on both factuality and relevance. While the system achieves consistent scores in citation-based evaluation, with strict F1 scores of 44.7 (micro) and 46.4 (macro), it underperforms in text generation metrics, particularly in BLEU (0.6) and BERTScore (23.9). This suggests that while the model effectively identifies relevant evidence, further refinement is required to enhance the fluency and linguistic alignment of generated responses.

Metric	Score	Micro	Macro
Overall Score	36.4	–	–
Factuality	44.7	–	–
Relevance	28.1	–	–
Strict Precision	–	58.7	63.6
Strict Recall	–	36.1	42.7
Strict F1	–	44.7	46.4
Lenient Precision	–	61.7	68.5
Lenient Recall	–	35.9	41.4
Lenient F1	–	45.4	47.8
BLEU	0.6	–	–
ROUGE-L	19.9	–	–
SARI	49.0	–	–
BERTScore	23.9	–	–
AlignScore	43.0	–	–
MEDCON (UMLS)	32.4	–	–

Table 1: Official evaluation results on the test dataset across overall performance, citation-based, and text generation metrics.

4.2 Ablation Study

Table 2 reports the lenient F1 scores across the four test cases for each individual model, all pairwise and three-way combinations, and the complete four model ensemble. Among the single models, Bio_ClinicalBERT performs best, which is consistent with its clinical-domain pretraining. However, the standard BERT model—despite lacking biomedical specialization—proves surprisingly effective, particularly in combination with other models. In fact, BERT appears to play a stabilizing role in most ensemble variants. Its inclusion consistently improves performance, often more than one might expect given its standalone score. This suggests that general-domain representations may provide complementary context cues that specialized models overlook especially when clinical language overlaps with common phrasing. Performance improves steadily as models are added, with all three-model combinations outperforming any two-model setup. Interestingly, the top three-model combination excludes BERT, but only slightly edges out the BERT-inclusive variants›. Ultimately, the full ensemble outperforms all others, confirming that diversity in model training is meaningful to relevance prediction.

4.3 Resource Usage

The entire pipeline—including preprocessing, relevance classification, and answer generation—runs comfortably within the 24GB VRAM limit of a single NVIDIA RTX 4090 GPU. During inference, the relevance classification stage takes approximately 1 second per case on average. The answer generation stage, which uses the Mistral Small 3.1 model

Variant	F1
BERT	0.524
Bio_ClinicalBERT	0.544
BlueBERT	0.515
MedEmbed	0.507
BERT + Bio_ClinicalBERT	0.579
BERT + BlueBERT	0.563
BERT + MedEmbed	0.522
Bio_ClinicalBERT + BlueBERT	0.532
Bio_ClinicalBERT + MedEmbed	0.552
BlueBERT + MedEmbed	0.546
BERT + Bio_ClinicalBERT + BlueBERT	0.563
BERT + Bio_ClinicalBERT + MedEmbed	0.602
BERT + BlueBERT + MedEmbed	0.602
Bio_ClinicalBERT + BlueBERT + MedEmbed	0.603
Full ensemble (all 4)	**0.619**

Table 2: Ablation study: F1 scores for each single model, model combination, and the full ensemble.

via the Ollama interface, averages 15 seconds per case. Post-processing, which involves formatting corrections and citation verification, adds an additional 0.001 seconds per case on average and is performed entirely on CPU.

Altogether, the full inference pipeline processes each case in about 16 seconds end-to-end. These performance characteristics confirm the system's suitability for real-time or near-real-time deployment in clinical or low-latency environments. Additionally, the total cost for running inference over the full test set is negligible when using standard compute infrastructure, making the approach both scalable and accessible.

4.4 Error Analysis

The model exhibits false positives in sentences with clinical terms or medication instructions that are not directly relevant to the query, such as "You were started on a milrinone drip, with improvement in your heart's pump function". This suggests over-reliance on clinical terminology rather than contextual alignment. Conversely, false negatives often involve broader prognostic statements or mental health assessments, where relevance is implied across multiple sentences. This indicates a need for improved contextual understanding to handle less explicit but clinically relevant content.

5 Conclusions

This paper presents a modular pipeline for relevance classification and grounded answer generation in the ArchEHR-QA shared task, operating under constrained computational resources. The use of pre-trained models with minimal fine-tuning proved effective in leveraging both general-domain and medical-specific knowledge, resulting in consistent citation-based evaluation scores. However, lower scores in BLEU and BERTScore indicate that further refinement is necessary to improve the fluency and linguistic alignment of generated responses. Future work will explore methods for enhancing response generation, including advanced grounding techniques and multi-sentence contextual modeling.

Limitations

The reliance on a single RTX 4090 GPU constrained the computational capacity available for training and fine-tuning, limiting the scope of model experimentation and hyperparameter optimization. Additionally, the development dataset consisted of only 20 cases, restricting the diversity of clinical scenarios encountered during training and potentially impacting the system's ability to generalize effectively.

Acknowledgements

The authors acknowledge the support of project PN 23 38 01 01, "Contributions to the consolidation of emerging technologies specific to the Internet of Things and complex systems," which provided technical resources essential to this research. Ana Uban was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS - UEFISCDI, project SIROLA, number PN-IV-P1-PCE-2023-1701, within PNCDI IV, and by CCCDI - UEFISCDI, project number PN-IV-P7-7.1-PTE-2024-0046, within PNCDI IV.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Abhinand Balachandran. 2024. Medembed: Medical-focused embedding models.

Jayetri Bardhan, Kirk Roberts, and Daisy Zhe Wang. 2023. Question answering for electronic health records: A scoping review of datasets and models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. MIMIC-IV-Note: Deidentified free-text clinical notes.

Alistair Johnson, Tom Pollard, and Roger Mark. 2023b. MIMIC-III clinical database.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries.

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.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Mistral AI. 2025. Mistral small 3.1. `https://mistral.ai/news/mistral-small-3-1`.

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.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Pablo Romero, Lifeng Han, and Goran Nenadic. 2025. Medication extraction and entity linking using stacked and voted ensembles on LLMs. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 303–315, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2024. Rationale-guided retrieval augmented generation for medical question answering. *Preprint*, arXiv:2411.00300.

Sarvesh Soni, Surabhi Datta, and Kirk Roberts. 2023. Quehry: A question answering system to query electronic health records. *Journal of the American Medical Informatics Association*, 30(6):1091–1102.

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.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

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. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Appendix A: Prompt Templates

Generation Prompt

You are a professional clinical assistant.

Using only the provided evidence, write a concise, clinical answer to the clinician's question. Do not include any citations. The answer must be no more than 75 words.

For example: **Clinician Question:** Why did they perform the emergency salvage repair on him?

Evidence: - |1| He was transferred to the hospital on 2025-1-20 for emergent repair of his ruptured thoracoabdominal aortic aneurysm. - |2| He was immediately taken to the operating room where he underwent an emergent salvage repair of ruptured thoracoabdominal aortic aneurysm with a 34-mm Dacron tube graft using deep hypothermic circulatory arrest. - |8| Thoracoabdominal wound healing well with exception of very small open area mid wound that is @1cm around and 1/2cm deep, no surrounding erythema.

Answer: His aortic aneurysm was caused by the rupture of a thoracoabdominal aortic aneurysm, which required emergent surgical intervention |1|. He underwent a complex salvage repair using a 34-mm Dacron tube graft and deep hypothermic circulatory arrest to address the rupture |2|. The extended recovery time and hospital stay were necessary due to the severity of the rupture and the complexity of the surgery, though his wound is now healing well with only a small open area noted |8|.

Another example: **Clinician Question:** Why was ERCP recommended over a medication-based treatment for CBD sludge?

Evidence: - |1| During the ERCP a pancreatic stent was required to facilitate access to the biliary system (removed at the end of the procedure), and a common bile duct stent was placed to allow drainage of the biliary obstruction caused by stones and sludge. - |5| On hospital day 4 (post-procedure day 3) the patient returned to ERCP for re-evaluation of her biliary stent as her LFTs and bilirubin continued an upward trend. - |6| On ERCP the previous biliary stent was noted to be acutely obstructed by biliary sludge and stones. - |7| As the patient's INR was normalized to 1.2, a sphincterotomy was safely performed, with removal of several biliary stones in addition to the common bile duct stent.

Answer: Medications can sometimes help in managing bile duct sludge, but in this case, ERCP was necessary due to the severity of the obstruction and its complications. The initial ERCP revealed significant biliary obstruction caused by sludge and stones, requiring the placement of a stent to restore bile drainage |1|. However, even after this intervention, the liver function tests and bilirubin levels continued to rise, indicating that the obstruction was not fully resolved |5|. A follow-up ERCP confirmed that the stent itself had become acutely obstructed by sludge and stones, necessitating further intervention |6|. During this procedure, a sphincterotomy was performed, and several stones were physically removed, which medications alone could not have achieved |7|. These findings confirm that ERCP was essential in addressing his condition and preventing further complications.

Now, using the provided evidence, write a concise, clinical answer to the clinician's question. Do not include any citations (this will be done in a future step). The answer must be no more than 75 words.

Question: {*clinician_question*}

Evidence: {*evidence_text*}

Retry Prompt for Length Enforcement

The last answer was too long ({*word_count*} words). Please try again, but make sure the answer is no more than {*MAX_WORDS*} words.

Last answer: {*last_answer*}

Grounding Prompt

You are a clinical documentation assistant.

You will be given: - A clinical question - A professional clinical answer to the question - A list of evidence sentences with sentence IDs in the format |sentence_id|

Your task is to add appropriate citations to the clinical answer. For each statement in the answer, cite the sentence(s) from the evidence that support it using the format |sentence_id|. If multiple sentences support a statement, separate them with commas, e.g., |3,4,7|. Do not use ranges like |1-3|. Cite only at the end of the sentence (after the period), and always add a newline after the citation. Do not add a newline after the final sentence. Do not change the wording of the answer. Simply append the appropriate citation(s).

Clinician Question: {*clinician_question*}

Clinical Answer: {*answer*}

Evidence: {*evidence_text*}