SzegedAI at ArchEHR-QA 2025: Combining LLMs with traditional methods for grounded question answering

Soma Bálint Nagy, Bálint Nyerges, Zsombor Mátyás Kispéter, Gábor Tóth, András Tamás Szlúka, Gábor Kőrösi, Zsolt Szántó, Richárd Farkas,

Institute of Informatics, University of Szeged
2. Árpád tér, Szeged, Hungary
{nagysoma,tothg,korosig,szantozs,rfarkas}@inf.u-szeged.hu

Abstract

In this paper, we present the SzegedAI team's submissions to the ArchEHR-QA 2025 shared task. Our approaches include multiple prompting techniques for large language models (LLMs), sentence similarity methods, and traditional feature engineering. We are aiming to explore both modern and traditional solutions to the task. To combine the strengths of these diverse methods, we employed different ensembling strategies.

1 Introduction

The ArchEHR-QA 2025 shared task (Soni and Demner-Fushman, 2025b) aimed to help reduce the workload of clinicians by automatically generating answers to patients' questions. These answers were based on information from patients' electronic health records (EHRs) (Soni and Demner-Fushman, 2025a). The goal was to ensure that the answers were grounded in the clinical notes, with clear references to the specific sentences in the records. The task focused on two main evaluation criteria: factuality, which checks if the references are correct, and relevancy, which evaluates the quality of the answers.

In our solution, we combined strategies based on large language models (LLMs) with classical NLP techniques, such as the bag-of-words representation of overlapping terms between the question and the sentences. Our results include a comparison of different LLMs, such as Gemini (Team et al., 2024), Gemma 3 (Team et al., 2025), LLama (Grattafiori et al., 2024) and its medical fine-tuned versions (Ankit Pal, 2024; Christophe et al., 2024; Kim et al., 2025). We applied prompting strategies that either directly generate answers with references or select relevant sentences and generate responses from them. Additionally, we combined the outputs of the models using a voting mechanism, along with feature-rich classification techniques trained on the development set.

2 System Overview

We developed two main approaches:

1. **Pipeline Approach**: A two-step process that first identifies essential sentences in the clinical notes and then generates an answer based on these sentences.

2. **End-to-End Approach**: A single-step process that directly generates responses with appropriate citations using an LLM.

Our primary focus was on the pipeline approach, where we experimented with different methods for both essential sentence identification and answer generation. For essential sentence identification, we looked at the problem from both classical machine learning and LLM-based perspectives. The ML approach utilized feature engineering with lexical and semantic similarity metrics between questions and clinical note sentences, and other textual features. While the LLM-based approaches employed various prompting strategies to identify essential sentences through direct citation, two-agent interaction, and pairwise question-sentence evaluation.

We also explored ensemble techniques for essential sentence identification that combined the strengths of our various approaches through voting mechanisms and feature-rich classification. These ensemble models incorporated predictions from previous methods to improve overall performance.

For answer generation in our pipeline approach, we developed methods that used the identified essential sentences as input to craft concise, coherent responses that answered the question while properly citing the source sentences.

In our end-to-end approach, we prompted LLMs with carefully designed instructions to simultaneously identify relevant clinical evidence and generate coherent answers with citations in a single step.

BioNLP 2025 Shared Tasks, pages 136–149
August 1, 2025 ©2025 Association for Computational Linguistics

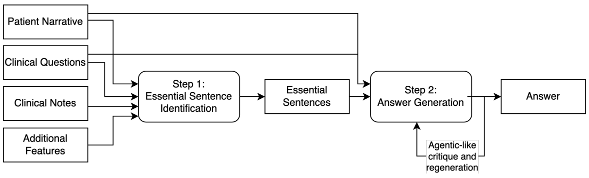
Patient Narrative
Clinical Questions
Clinical Notes
Additional Features
Step 1: Essential Sentence Identification
Essential Sentences
Step 2: Answer Generation
Answer
Agentic-like critique and regeneration

Figure 1: Overview of our pipeline system architecture for the ArchEHR-QA task, showing the two-step process of essential sentence identification followed by answer generation.

Both approaches were enhanced with an agentic reflection loop where initial responses were programmatically validated against task requirements (citation format, answer length, coverage of essential information) and iteratively refined based on specific feedback.

3 Methods

Sections 3.1 and 3.2 detail the various techniques we tested for the pipeline approach, while section 3.3 summarizes our procedure for the end-to-end approach.

3.1 Essential Sentence Identification

The first step of our pipeline approach was the identification of the essential sentences. We treated this problem as a binary classification task (essential or not-relevant), without considering "supplementary" as a separate category. We explored three main approaches:

3.1.1 Supervised Machine Learning-based Classification

We implemented a traditional machine learning approach using a LightGBM classifier (Ke et al., 2017), treating each sentence as a separate training or test instance. The following feature templates were used:

- Bag-of-words representations and overlap between question and sentence
- Semantic embeddings and cosine similarities (between question-sentence and between adjacent sentences)
- Length features (question, sentence, and their difference)
- Positional indicators (first/last sentence in note)

3.1.2 LLM-based Classification

We leveraged the contextual awareness and potential domain knowledge of LLMs through various prompting strategies (the prompts are available in Appendix A):

Answer with References This approach prompted the LLM to generate answers with citations to relevant clinical note sentences, which were labeled as essential. Unlike our pipeline's answer generation step, it omitted validation for length limits and formatting requirements. We also developed a two-stage variation (v2) that first identified the key sentence answering the question, then found supporting context sentences.

Agentic This method used two LLM instances: one generated an uncited answer, while the second identified supporting sentences from the clinical notes, which were labeled as essential.

References Only This approach focused solely on identifying essential sentences without generating a complete answer. The LLM was prompted to analyze the question and clinical notes, then output the numbers of sentences containing essential information. We used chain-of-thought reasoning and tested both zero-shot and one-shot variants.

Question-Sentence Compare This strategy evaluated individual question-sentence pairs rather than full cases, with the LLM classifying each sentence as essential or not. For reliability, we applied majority voting across three separate evaluations of each sentence.

3.1.3 Ensemblers

We developed two distinct ensemble approaches for essential sentence identification:

Supervised Ensembler This approach combined traditional machine learning features with the predictions from our various LLM-based methods as additional input features. This hybrid method leveraged both the structured learning of traditional classifiers and the contextual understanding provided by LLMs.

Answer with references - voting We created five variations of our "Answer with References" prompt with slight modifications. Sentences that were marked as essential by at least three of the five generated answers were considered essential in the final output, creating a majority-voting ensemble.

3.2 Answer Generation

The second step of our pipeline approach is the answer generation. Here our prompts contained the patient narrative, clinician question, and the full list of sentences identified as essential by our classification methods. We developed an iterative prompting strategy with an agentic reflection loop where each generated answer was programmatically validated against several key requirements: proper citation formatting, answer length constraints, comprehensive coverage of all essential information, and proper citation of all identified essential sentences.

When an answer failed to meet any of these criteria, we provided the LLM with the original prompt, the unsatisfactory answer, and specific feedback identifying the shortcomings. This initiated an iterative refinement process where the model would revise its response based on the targeted feedback, continuing until all quality requirements were satisfied.

3.3 End-to-End Approach

In contrast to our pipeline approach, we also explored an end-to-end method that directly generated answers with appropriate citations in a single step. For this approach, we provided the LLM with all sentences from the clinical notes rather than prefiltering for essential ones. The prompt explicitly specified that not all sentences contained relevant information and that the model should only cite sentences that directly underpinned its answer.

The end-to-end prompts instructed the model to generate a coherent answer using the clinical notes, include proper citations, address key aspects of the question concisely, and adhere to formatting requirements—all in a single step.

This approach was also enhanced with an agentic reflection loop, though with a different set of validation criteria. Since no separate sentence identification step existed, validation focused primarily on formatting correctness, citation syntax, and answer length constraints.

4 Results

In this section, we present the results of our methods. We begin by showing the performance of our models on the development set, followed by the performance of our submissions on the test set.

4.1 Experimental setup

On the development set, we focus on factuality (essential sentence identification) as the primary criteria.

The supervised machine learning-based classifier was trained on the development set with 100 estimators, gradient boosting decision trees, a fixed random seed of 42, and a minimum of 10 data points in each leaf. The model was validated using k-fold cross-validation, where k was 5. To calculate semantic representation we used LaBSE (Feng et al., 2020).

In our experiments, we compared various LLMs to find the best for the shared task[1]. Besides our baseline models, LLama 3.3 70B and Gemma 3 27B, we utilized fine-tuned models for different biomedical goals. Llama3-OpenBioLLM-70B model fine-tuned for biomedical tasks using DPO and a curated medical instruction dataset. Llama3-Med42-70B is optimized for medical question answering and clinical knowledge with instruction tuning. Llama-3-Meerkat-70B (Kim et al., 2025) is built for medical reasoning, trained with synthetic CoT data and diverse instruction datasets. Along with the open source models, we also used Gemini 1.5 Flash model.

4.2 Essential Sentence Identification

First, we evaluated our systems on the development set, which is shown in the Table 1.

Supervised classification Despite the limited number of training examples, our supervised machine learning-based classification model that mainly applies bag-of-words and semantic similarity-based features performed comparably to many prompt-based solutions. It achieved better results than 9 out of 13 LLM-based approaches.

LLMs Among the tested LLMs, the Gemini 1.5 Flash outperformed both the original and biomedical LLaMA 70Bs and Gemma 3 27B by a large margin. In the challenge of 70B LLaMa variants, 2 out of 3 fine-tuned models preceded the original model, where the Llama-3-Meerkat-70B was the best. Interestingly, the smaller Gemma model, which was not fine-tuned on medical data, achieved comparable results to the best LLaMA model.

Prompting strategies When comparing prompting strategies, the best results were obtained

[1] We used 4 A100 GPU for the open sourced LLMs.

	LLM	strict-micro			strict-macro		
		P	R	F1	P	R	F1
Competition baseline	Llama 3.3 70B	0.634	0.326	0.431	**0.703**	0.471	0.494
Supervised classifier	-	0.521	0.529	0.525	0.510	0.514	0.499
Answer with references	Gemini	0.566	0.558	0.562	0.608	**0.638**	0.578
Answer with references	Llama 3.3 70B	0.397	0.362	0.379	0.416	0.349	0.357
Answer with references	Llama3-Med42-70B	0.341	0.341	0.341	0.358	0.400	0.340
Answer with references	Llama3-OpenBioLLM-70B	0.333	0.275	0.406	0.309	0.308	0.289
Answer with references	Llama-3-Meerkat-70B	0.336	0.406	0.385	0.360	0.434	0.362
Answer with references	Gemma 27B	0.400	0.406	0.403	0.419	0.428	0.398
Answer with references v2	Gemini	0.631	0.384	0.477	0.651	0.443	0.477
Agentic	Gemini	0.500	0.442	0.469	0.583	0.530	0.495
References only - zero shot	Gemini	0.657	0.500	0.568	0.659	0.568	0.574
References only - 1 shot	Gemini	0.699	0.522	0.598	0.662	0.591	0.583
Question - sentence compare	Gemini	0.477	0.536	0.505	0.481	0.519	0.457
Question - sentence compare	Gemini	0.503	0.536	0.519	0.517	0.518	0.462
End-to-end	Gemini	0.693	0.507	0.587	0.534	0.438	0.473
Answer with references - voting	Gemini	0.514	0.398	0.449	0.538	0.490	0.454
Supervised ensembler	Gemini	**0.750**	**0.608**	**0.672**	0.685	0.586	**0.616**

Table 1: Factuality results of the independent systems on the development set. The `Competition baseline` used the LLaMA 3.3 70B model in a zero-shot setting prompting it to generate cited answers; if responses were invalid, they retried up to five times to get a valid one. Detailed descriptions of the `Supervised classifier` method can be found in Section 3.1.1; `Answer with references (V2)`, `Agentic`, `References only` and `Question - sentence compare` are in Section 3.1.2; `End-to-end` in 3.3; and `Answer with references - voting` and `Supervised ensembler` are in 3.1.3.

with the `References only` and `Answer with references` approaches for sentence identification, but the `End-to-end` approch also acheived similarly high score.

Ensemblers The voting method over the `Answer with references` can't improve the performance. Instead of the `Supervised ensembler` that applies all of the Gemini-based system's output as features besides the features of the `Supervised classifier`, achieved the highest score on the development set.

	SC	E2E	SE
Overall	0.321	0.407	**0.427**
Overall Factuality	0.317	0.470	**0.472**
Strict F1 (micro)	0.317	0.470	**0.472**
Strict F1 (macro)	0.309	**0.523**	0.514
Overall Relevance	0.325	0.344	**0.382**
BLEU	0.018	0.008	**0.032**
ROUGELsum	0.227	0.211	**0.292**
SARI	0.558	0.597	**0.642**
BERTScore	**0.288**	0.275	0.191
AlignScore	0.272	**0.631**	0.195
MEDCON (UMLS)	**0.586**	0.344	0.278

Table 2: Official scores of our systems on the test set.

4.3 Submissions

We selected three distinct models as submissions to reflect the variety of approaches we had previously evaluated on the development set, results presented in the Table 2. The first model, `Supervised classifier (SC)`, aimed to evaluate the performance of traditional machine learning methods on the shared task. The `End-to-end (E2E)` model was one of the most purely prompt-based solutions, and we uploaded our best system from the development set, the `Supervised ensembler (SE)`.

The `SC` model performed notably worse on the test set than on the development set. Since we did not use the development set for hyperparameter tuning during cross-validation, we suspect that the limited amount of training data failed to generalize well to the test set. A similar pattern was observed with our `SE` model. But in this case, the factuality score is matched with the `E2E` model, which was in third place on the development set. In the case of relevance, the `SE` model, which generates answers based on selected essential sentences, outperformed the `E2E` model. Consequently, the `SE` also achieved a higher score on the overall metric, so we selected this model as our official submission.

5 Conclusion

In this paper, we presented the SzegedAI team's submissions to the ArchEHR-QA 2025 shared task. Our models combined traditional machine learning techniques with LLM-based predictions. We explored a range of models and prompting strategies, and integrated their outputs using a feature-rich classification framework to identify the most relevant information from clinical notes in response to patient questions. Our submission achieved 11th place in the automatic evaluation of the shared task.

Limitations

In this paper, we relied heavily on the development set for evaluations, but the small size of this dataset limits the accurate comparison of the different methods.

Most of our LLM-based methods were limited to one prompt per question, except the `Agentic`, `End-to-end`, and `Answer generation` methods, which were limited to five cycles, and the `Question-sentence compare` applied an LLM call for each sentence in a clinical note.

While our supervised machine learning-based systems performed well on the development set, their performance dropped on the test set, suggesting potential overfitting and limited generalization due to the small training size. Increasing the amount of training data would likely improve results, but the `Supervised classifier` is inherently less generalizable than LLMs.

Our evaluation focused on factuality metrics, with less emphasis on the relevance of the answer, which plays a critical role in real-life applications.

Acknowledgments

The research is supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory.

We acknowledge the Digital Government Development and Project Management Ltd. for awarding us access to the Komondor HPC facility based in Hungary.

References

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, Thomas Fetherston, Donghee Choi, Soo Heon Kwak, Qingyu Chen, and Jaewoo Kang. 2025. Small language models learn enhanced reasoning skills from medical textbooks. *npj Digital Medicine*, 8(1):240.

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.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

A Prompts for essential sentence identification

This section shows the prompts that were applied to the results of the paper.

A.1 Prompt for "Answer with References"

```
Task: Generate a concise, helpful answer to a patient's health question using only information from the clinical note. Each statement in your answer must be grounded in specific sentences from the note.
Example:
Patient's Narrative: Took my 59 yo father to ER ultrasound discovered he had an aortic aneurysm. He had a salvage repair (tube graft). Long surgery / recovery for couple hours then removed packs. why did they do this surgery????? After this time he spent 1 month in hospital now sent home.
Patient's Question: why did they do this surgery?????
Clinician's Rephrased Question: Why did they perform the emergency salvage repair on him?
Clinical Note (numbered sentences):
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.
3: Please see operative note for details which included cardiac arrest x2.
4: Postoperatively he was taken to the intensive care unit for monitoring with an open chest.
5: He remained intubated and sedated on pressors and inotropes.
6: On 2025-1-22, he returned to the operating room where he underwent exploration and chest closure.
7: On 1-25 he returned to the OR for abd closure JP/ drain placement/ feeding jejunostomy placed at that time for nutritional support.
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.
9: Packed with dry gauze and covered w/DSD.
Example 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).
Now, please generate an answer for the following case:
Patient's Narrative: {patient_narrative}
Patient's Question: {patient_question}
Clinician's Rephrased Question: {clinician_question}
Clinical Note (numbered sentences):
{numbered_note}
Instructions:

First, carefully identify which sentences are ESSENTIAL to answering the clinician's rephrased question. Focus on sentences that directly explain the medical reasoning, procedures performed, and clinical findings.
When writing your answer, ONLY include information from these essential sentences. Each statement in your answer MUST be supported by at least one citation.
For each statement in your answer, cite the specific sentence number(s) that support it using parentheses, e.g., "The procedure was successful (3, 5)."
Be very precise with your citations - only cite sentences that directly support each specific claim you make.

Your Answer:
```

A.2 Prompts "Agentic LLM classification"

A.2.1 Stage 1: Answer generation prompt

```
Task: Generate a helpful, concise answer to a patient's health question using only information from the clinical note.
Example:
Patient's Narrative: Took my 59 yo father to ER ultrasound discovered he had an aortic aneurysm. He had a salvage repair (tube graft). Long surgery / recovery for couple hours then removed packs. why did they do this surgery????? After this time he spent 1 month in hospital now sent home.
Patient's Question: why did they do this surgery?????
Clinician's Rephrased Question: Why did they perform the emergency salvage repair on him?
Clinical Note (numbered sentences):
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.
3: Please see operative note for details which included cardiac arrest x2.
4: Postoperatively he was taken to the intensive care unit for monitoring with an open chest.
5: He remained intubated and sedated on pressors and inotropes.
6: On 2025-1-22, he returned to the operating room where he underwent exploration and chest closure.
7: On 1-25 he returned to the OR for abd closure JP/ drain placement/ feeding jejunostomy placed at that time for nutritional support.
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.
9: Packed with dry gauze and covered w/DSD.
Example Answer:
His aortic aneurysm was caused by the rupture of a thoracoabdominal aortic aneurysm, which required emergent surgical intervention. He underwent a complex salvage repair using a 34-mm Dacron tube graft and deep hypothermic circulatory arrest to address the rupture. 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.
Now, please generate an answer for the following case:
Patient's Narrative: {patient_narrative}
Patient's Question: {patient_question}
Clinician's Rephrased Question: {clinician_question}
Clinical Note (numbered sentences):
```

```
{numbered_note}
Instructions:

Answer the clinician's rephrased question directly and clearly.
Use only information found in the clinical note.

Your Answer:
```

A.2.2 Stage 2: Source identification prompt

```
Task: Identify which sentences from the clinical note support statements in the patient answer.

Example:
Clinical Note (numbered sentences):
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.
3: Please see operative note for details which included cardiac arrest x2.
4: Postoperatively he was taken to the intensive care unit for monitoring with an open chest.
5: He remained intubated and sedated on pressors and inotropes.
6: On 2025-1-22, he returned to the operating room where he underwent exploration and chest closure.
7: On 1-25 he returned to the OR for abd closure JP/ drain placement/ feeding jejunostomy placed at that time for nutritional
↪  support.
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.
9: Packed with dry gauze and covered w/DSD.
Example input text:
His aortic aneurysm was caused by the rupture of a thoracoabdominal aortic aneurysm, which required emergent surgical
↪  intervention. He underwent a complex salvage repair using a 34-mm Dacron tube graft and deep hypothermic circulatory
↪  arrest to address the rupture. 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.
Essential Sentences: 1, 2, 8

Now, please generate an answer for the following case:
Clinical Note (numbered sentences):
{numbered_note}
Input text:
{generated_answer}
Instructions:

Carefully analyze the answer and identify ALL sentences from the clinical note that directly support information in the answer.
Do not include sentences that contain information not referenced in the text.
List ONLY the sentence numbers (without any additional text) in a comma-separated format.
Your response should follow this format exactly:

Essential Sentences: [list of numbers]
For example: "Essential Sentences: 1, 3, 5, 7"
```

A.3 Prompts for "Answer with references v2"

A.3.1 Stage 1: Best sentence identification prompt

```
Task: Identify the SINGLE BEST sentence from the clinical note that directly answers the clinician's question.

Patient's Narrative: {patient_narrative}
Patient's Question: {patient_question}
Clinician's Rephrased Question: {clinician_question}

Clinical Note (numbered sentences):
{numbered_note}

Instructions:

    Analyze each sentence in the clinical note carefully.
    Identify the ONE sentence that most directly answers the clinician's question about why a procedure was performed, what
    ↪  caused a condition, how something was treated, or other clinical reasoning.
    Choose the sentence that contains the core explanation, not just related information.
    Provide ONLY the sentence number in your response, with no additional text.

Few-Shot Examples:

Example 1:
Patient Question: "My question is if the sludge was there does not the medication help in flushing it out? Whether ERCP was
↪  the only cure?"
Clinician Question: "Why was ERCP recommended to him over continuing a medication-based treatment?"
Clinical Note:

    Brief Hospital Course:
    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.
    However, due to the patient's elevated INR, no sphincterotomy or stone removal was performed.
    Frank pus was noted to be draining from the common bile duct, and post-ERCP it was recommended that the patient remain on
    ↪  IV Zosyn for at least a week.
    The Vancomycin was discontinued.
    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.
    On ERCP the previous biliary stent was noted to be acutely obstructed by biliary sludge and stones.
    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.
    At the conclusion of the procedure, retrograde cholangiogram was negative for filling defects.

Best Sentence Answer: 2

Reasoning: Sentence 2 is the best single sentence because it directly explains why ERCP was necessary - it reveals that stones and sludge were causing a biliary obstruction that required stent placement to allow drainage. This is the core reason why medication alone wouldn't be sufficient - there was a physical blockage that needed mechanical intervention.

Example 2:
Patient Question: "I overdosed October 4th on trihexyphenidyl, thorazine, and cocaine. I have had chest pain in my left upper quadrant ever since. Any ideas?"
Clinician Question: "Is the pain connected to the overdose or something else?"
Clinical Note:

    Brief Hospital Course:
    Bipolar d/o, PTSD, schizophrenia: Psychiatry consult recommended that all psych medications be held until they could be re-prescribed by pt's outpatient psychiatrist.
    During hospital course, thorazine was restarted but discontinued soon after because pt became tachycardic; pt remained asymptomatic during these episodes of tachycardia.
    Tachycardia resolved with discontinuation of thorazine, IV hydration, and small dose of IV benzodiazepene x 1.
    Social work consult was obtained because pt did not have a PCP nor did he have a psychiatrist.
    He could not see his former psychiatrist due to insurance reasons.
    With the help of social work, pt was set up with a PCP who would be able to refer him to a new psychiatrist in a timely fashion.
    He was instructed to follow-up with his new psychiatrist to restart his psychiatric medications.
    Chest pain: Pt complained of chest pain during hospital course that appeared musculoskeletal as it was reproducible with palpation and pt reported more pain with movement.
    EKG showed no ischemic changes and troponins were flat x 4. CK was elevated, peaking at 1405 but downtrended without any intervention.
    TTE was obtained due to history of cocaine use to rule out cardiac events.
    EF was >55%; TTE was unremarkable.
    He was monitored on telemetry without significant events.
    Discharge Instructions:
    It was a pleasure taking care of you at the hospital.
    You were admitted with confusion that was likely due to a combination of the medications you were taking and the street drugs that you may have also been used.
    Your heart rhythm was monitored because many of these drugs can affect your heart.
    Your EKG and blood tests showed that you likely did not have a heart attack.
    An ultrasound of your heart was also normal.
    Your confusion cleared during your hospital stay.
    You were seen by our psychiatry team who recommended holding all of your medications while you were in the hospital.
    It is very important that you follow-up with a primary care doctor who can refer you to a psychiatrist.
    This psychiatrist can then prescribe to you the medications you were normally taking.

Best Sentence Answer: 9

Reasoning: Sentence 9 is the best choice because it directly addresses the nature of the chest pain, identifying it as musculoskeletal based on clinical examination (reproducible with palpation and worsening with movement). This directly answers whether the pain is connected to the overdose or something else by suggesting a musculoskeletal cause.

Your Answer:
```

A.3.2 Stage 2: Context sentences identification prompt

```
Task: Identify additional sentences from the clinical note that provide necessary context for understanding the answer to the clinician's question.

Patient's Narrative: {patient_narrative}
Patient's Question: {patient_question}
Clinician's Rephrased Question: {clinician_question}

Clinical Note (numbered sentences):
{numbered_note}

The MAIN sentence that answers the question is:
Sentence {best_sentence_num}: {best_sentence_text}

Instructions:

    Analyze the clinical note to identify any OTHER sentences that provide necessary context to fully understand the answer.
    Include sentences that:
        Explain medical terminology used in the main answer
        Provide evidence supporting the main answer
        Show treatment outcomes that validate the answer
        Describe test results that confirm the diagnosis or treatment decision
        Explain why alternative treatments were not chosen
    Exclude sentences that:
        Repeat information already in the main sentence
        Contain general information not directly related to the question
        Focus on administrative details rather than clinical reasoning
    List ONLY the sentence numbers in your response, separated by commas (e.g., "3, 5, 9").
    If no additional context sentences are needed, respond with "None".
    Limit your selection to the most relevant sentences (typically 2-5 sentences).

Few-Shot Examples:

Example 1:
Patient Question: "My question is if the sludge was there does not the medication help in flushing it out? Whether ERCP was the only cure?"
Clinician Question: "Why was ERCP recommended to him over continuing a medication-based treatment?"
Clinical Note:

    Brief Hospital Course:
    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.
    However, due to the patient's elevated INR, no sphincterotomy or stone removal was performed.
    Frank pus was noted to be draining from the common bile duct, and post-ERCP it was recommended that the patient remain on IV Zosyn for at least a week.
    The Vancomycin was discontinued.
    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.
    On ERCP the previous biliary stent was noted to be acutely obstructed by biliary sludge and stones.
    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.
    At the conclusion of the procedure, retrograde cholangiogram was negative for filling defects.

Main sentence that answers the question is:
Sentence 2: 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.

Context Sentences Answer: 6, 7, 8

Reasoning for including these context sentences:

    Sentence 6 shows that even after initial treatment, the patient's liver function tests continued to worsen, indicating that medication alone was not sufficient
    Sentence 7 demonstrates that the biliary stent became obstructed again by sludge and stones, further proving that physical removal was necessary
    Sentence 8 shows that once conditions allowed (normalized INR), a sphincterotomy was performed to physically remove the stones, which medication alone couldn't accomplish

Reasoning for NOT including other potential sentences:

    Sentence 3 mentions elevated INR preventing sphincterotomy, but doesn't directly address why medication wouldn't work
    Sentence 4 mentions pus and antibiotics, which is related to infection treatment but not directly about sludge removal
    Sentence 9 only provides procedural outcome information without explaining why ERCP was necessary over medication

Example 2:
Patient Question: "I overdosed October 4th on trihexyphenidyl, thorazine, and cocaine. I have had chest pain in my left upper quadrant ever since. Any ideas?"
Clinician Question: "Is the pain connected to the overdose or something else?"
Clinical Note:

    Brief Hospital Course:
    Bipolar d/o, PTSD, schizophrenia: Psychiatry consult recommended that all psych medications be held until they could be re-prescribed by pt's outpatient psychiatrist.
    During hospital course, thorazine was restarted but discontinued soon after because pt became tachycardic; pt remained asymptomatic during these episodes of tachycardia.
    Tachycardia resolved with discontinuation of thorazine, IV hydration, and small dose of IV benzodiazepene x 1.
    Social work consult was obtained because pt did not have a PCP nor did he have a psychiatrist.
    He could not see his former psychiatrist due to insurance reasons.
    With the help of social work, pt was set up with a PCP who would be able to refer him to a new psychiatrist in a timely fashion.
    He was instructed to follow-up with his new psychiatrist to restart his psychiatric medications.
    Chest pain: Pt complained of chest pain during hospital course that appeared musculoskeletal as it was reproducible with palpation and pt reported more pain with movement.
    EKG showed no ischemic changes and troponins were flat x 4. CK was elevated, peaking at 1405 but downtrended without any intervention.
    TTE was obtained due to history of cocaine use to rule out cardiac events.
    EF was >55%; TTE was unremarkable.
    He was monitored on telemetry without significant events.
    Discharge Instructions:
    It was a pleasure taking care of you at the hospital.
    You were admitted with confusion that was likely due to a combination of the medications you were taking and the street drugs that you may have also been used.
    Your heart rhythm was monitored because many of these drugs can affect your heart.
    Your EKG and blood tests showed that you likely did not have a heart attack.
    An ultrasound of your heart was also normal.
    Your confusion cleared during your hospital stay.
    You were seen by our psychiatry team who recommended holding all of your medications while you were in the hospital.
    It is very important that you follow-up with a primary care doctor who can refer you to a psychiatrist.
    This psychiatrist can then prescribe to you the medications you were normally taking.

Main sentence that answers the question is:
Sentence 9: # Chest pain: Pt complained of chest pain during hospital course that appeared musculoskeletal as it was reproducible with palpation and pt reported more pain with movement.
```

```

Context Sentences Answer: 3, 10, 11, 12, 13

Reasoning for including these context sentences:

    Sentence 3 provides information about the thorazine (one of the overdosed medications) causing tachycardia, which could be related to the chest discomfort
    Sentence 10 rules out cardiac ischemia through EKG and troponin tests, while noting elevated CK (which can indicate muscle damage)
    Sentence 11 mentions additional cardiac testing due to history of cocaine use
    Sentence 12 shows normal heart function on ultrasound
    Sentence 13 confirms no cardiac events were detected during monitoring

Reasoning for NOT including other potential sentences:

    Sentences 16-19 from the discharge instructions contain similar information to sentences 10-13 but are written for the patient rather than providing additional clinical details
    Sentence 2 discusses psychiatric management but doesn't address the chest pain question
    Sentences 4-8 focus on medication management and discharge planning rather than explaining the chest pain

Your Answer:
```

A.4 Prompts for "References Only"

In this prompt, the question and the clinical note are given in a user prompt.

The system prompt:

```
Your task is to find essential sentences in a clinical note to answer a clinical question.
The clinical notes contain the history of a patient and details of a clinical event, you can select sentences from each category if neccessary.
There are always at least 3 essential sentences in the clinical note.
Try to find all of the relevant sentences in the clinical note to answer the question.

You can think step by step,
step 1: Analyze the question and the clinical note.
step 2: Find the essential sentences in the clinical note to answer the question. Write the reason why each sentence is essential or not.
step 3: To a separated last line list the ids of the essential sentences in the clinical note, in the following format:
1, 2, 3{example if example else ""}
```

The user prompt:

```
# Question
{patient_narrative}

# Clinical note
{json.dumps(clinical_note, indent=2)}
```

A.5 Prompt for "Question-Sentence Compare"

```
You are a medical expert. You will be given a question relating to a patient and a sentence which may or may not contain relevant information to answering the question. Your job is to tell wether the information is relevant or not-relevant.
This is the question of the patient:
{narrative}
{clinical_question}

The sentence is:
{sentence}

Does the sentence contain relevant information? Think carefully before you answer and end your answer with a definitive yes or no answer:
```

A.6 Prompts for "Answer with references - voting"

A.6.1 Prompt variation 1

```
Task: Generate a concise, helpful answer to a patient's health question using only information from the clinical note. Each statement in your answer must be grounded in specific sentences from the note.

Example:
Patient's Narrative: Took my 59 yo father to ER ultrasound discovered he had an aortic aneurysm. He had a salvage repair (tube graft). Long surgery / recovery for couple hours then removed packs. why did they do this surgery????? After this time he spent 1 month in hospital now sent home.

Patient's Question: why did they do this surgery?????

Clinician's Rephrased Question: Why did they perform the emergency salvage repair on him?
```

```

Clinical Note (numbered sentences):
**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.
**3:** Please see operative note for details which included cardiac arrest x2.
**4:** Postoperatively he was taken to the intensive care unit for monitoring with an open chest.
**5:** He remained intubated and sedated on pressors and inotropes.
**6:** On 2025-1-22, he returned to the operating room where he underwent exploration and chest closure.
**7:** On 1-25 he returned to the OR for abd closure JP/ drain placement/ feeding jejunostomy placed at that time for nutritional support.
**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.
**9:** Packed with dry gauze and covered w/DSD.

Example 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).

Now, please generate an answer for the following case:

Patient's Narrative: {patient_narrative}

Patient's Question: {patient_question}

Clinician's Rephrased Question: {clinician_question}

Clinical Note (numbered sentences):
{numbered_note}

Instructions:
1. First, carefully identify which sentences are ESSENTIAL to answering the clinician's rephrased question. Focus on sentences that directly explain the medical reasoning, procedures performed, and clinical findings.

2. When writing your answer, ONLY include information from these essential sentences. Each statement in your answer MUST be supported by at least one citation.

3. For each statement in your answer, cite the specific sentence number(s) that support it using parentheses, e.g., "The procedure was successful (3, 5)."

4. Be very precise with your citations - only cite sentences that directly support each specific claim you make.

Your Answer:
```

A.6.2 Prompt variation 2

```
Task: Answer a medical question based solely on the provided clinical note. Cite sentence numbers for each claim.

Example:
Patient's Narrative: Took my 59 yo father to ER ultrasound discovered he had an aortic aneurysm. He had a salvage repair (tube graft). Long surgery / recovery for couple hours then removed packs. why did they do this surgery????? After this time he spent 1 month in hospital now sent home.

Patient's Question: why did they do this surgery?????

Clinician's Rephrased Question: Why did they perform the emergency salvage repair on him?

Clinical Note (numbered sentences):
**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.
**3:** Please see operative note for details which included cardiac arrest x2.
**4:** Postoperatively he was taken to the intensive care unit for monitoring with an open chest.
**5:** He remained intubated and sedated on pressors and inotropes.
**6:** On 2025-1-22, he returned to the operating room where he underwent exploration and chest closure.
**7:** On 1-25 he returned to the OR for abd closure JP/ drain placement/ feeding jejunostomy placed at that time for nutritional support.
**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.
**9:** Packed with dry gauze and covered w/DSD.

Example 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).

Now answer this question:

Question: {clinician_question}

Clinical Note:
{numbered_note}

```

```
Instructions:
- Only use information directly from the note
- Each claim must have a citation in parentheses (e.g., "The surgery was successful (3)")
- Be concise and precise
- Only cite the most relevant sentences that directly answer the question

Your Answer:
```

A.6.3 Prompt variation 3

```
Task: Help a patient understand their medical situation by answering their question using information from their clinical note.

Example:
Patient's Narrative: Took my 59 yo father to ER ultrasound discovered he had an aortic aneurysm. He had a salvage repair (tube graft). Long surgery / recovery for couple hours then removed packs. why did they do this surgery????? After this time he spent 1 month in hospital now sent home.

Patient's Question: why did they do this surgery?????

Clinician's Rephrased Question: Why did they perform the emergency salvage repair on him?

Clinical Note (numbered sentences):
**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.
**3:** Please see operative note for details which included cardiac arrest x2.
**4:** Postoperatively he was taken to the intensive care unit for monitoring with an open chest.
**5:** He remained intubated and sedated on pressors and inotropes.
**6:** On 2025-1-22, he returned to the operating room where he underwent exploration and chest closure.
**7:** On 1-25 he returned to the OR for abd closure JP/ drain placement/ feeding jejunostomy placed at that time for nutritional support.
**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.
**9:** Packed with dry gauze and covered w/DSD.

Example 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).

Patient's Question: {patient_question}

Clinical Note:
{numbered_note}

Instructions:
1. Analyze which sentences in the note directly address the patient's question
2. Write a clear, concise answer citing only the most important sentences
3. Each statement must include sentence numbers in parentheses: (1) or (2, 3)
4. Be factual and only use information from the note

Your Answer:
```

A.6.4 Prompt variation 4

```
Task: Perform a structured medical note analysis to answer a clinical question.

Example:
Patient's Narrative: Took my 59 yo father to ER ultrasound discovered he had an aortic aneurysm. He had a salvage repair (tube graft). Long surgery / recovery for couple hours then removed packs. why did they do this surgery????? After this time he spent 1 month in hospital now sent home.

Patient's Question: why did they do this surgery?????

Clinician's Rephrased Question: Why did they perform the emergency salvage repair on him?

Clinical Note (numbered sentences):
**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.
**3:** Please see operative note for details which included cardiac arrest x2.
**4:** Postoperatively he was taken to the intensive care unit for monitoring with an open chest.
**5:** He remained intubated and sedated on pressors and inotropes.
**6:** On 2025-1-22, he returned to the operating room where he underwent exploration and chest closure.
**7:** On 1-25 he returned to the OR for abd closure JP/ drain placement/ feeding jejunostomy placed at that time for nutritional support.
**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.
**9:** Packed with dry gauze and covered w/DSD.

Example 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).

Clinical Question: {clinician_question}
Patient Context: {patient_narrative}

Clinical Note:
{numbered_note}

Process:
1. First, identify the 3-5 most relevant sentences that directly answer the question
2. Organize these sentences into a logical flow
3. Write a concise answer citing each sentence number in parentheses
4. Only include information that is explicitly stated in the cited sentences

Your Answer:
```

A.6.5 Prompt variation 5

```
Task: Use step-by-step reasoning to determine which sentences in a clinical note are essential to answering a medical question.

Example:
Patient's Narrative: Took my 59 yo father to ER ultrasound discovered he had an aortic aneurysm. He had a salvage repair (tube graft). Long surgery / recovery for couple hours then removed packs. why did they do this surgery????? After this time he spent 1 month in hospital now sent home.

Patient's Question: why did they do this surgery?????

Clinician's Rephrased Question: Why did they perform the emergency salvage repair on him?

Clinical Note (numbered sentences):
**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.
**3:** Please see operative note for details which included cardiac arrest x2.
**4:** Postoperatively he was taken to the intensive care unit for monitoring with an open chest.
**5:** He remained intubated and sedated on pressors and inotropes.
**6:** On 2025-1-22, he returned to the operating room where he underwent exploration and chest closure.
**7:** On 1-25 he returned to the OR for abd closure JP/ drain placement/ feeding jejunostomy placed at that time for nutritional support.
**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.
**9:** Packed with dry gauze and covered w/DSD.

Example 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).

Question to answer: {clinician_question}
Patient's original query: {patient_question}

Clinical Note:
{numbered_note}

Instructions:
1. First, break down what information is needed to answer the question
2. Identify only the sentences that contain this essential information
3. Write a concise answer using only these sentences
4. Include sentence numbers in parentheses after each claim: (1) or (2, 3)
5. Be precise - only cite sentences that directly support your statements

Your Answer:
```

B Answer generation prompt

```
# Medical question answering based on essential sentences

## Patient Information
**Patient narrative:** {patient_narrative}

**Clinician question:** {clinician_question}

## Numbered essential sentences from the clinical note
{essential_text}

## Task Instructions
1. Generate a mostly extractive response from the listed sentences, which serves as an answer for the question. You must maximize lexical overlap between the source sentences and the response, while providing a useful answer.
2. Each essential sentence must be cited at least once in your answer. Include the sentence numbers in parentheses after statements that use information from those sentences, e.g., (2) or (1, 3). Cite multiple sources separated by comma, when neccessary.
```

```
3. Citations must be at the end of each generated sentence.
4. Limit your answer to a maximum of {words_limit} words, but more than 50 words. (About 4-5 sentences.)

Be straight to the point with your answer to the question, avoid phrases like "Based on the sentences". Remember, you must maximize the similarity in the wording to the original sentences.

## [For Iteration i > 1] Previous Attempts
### Attempt {i-1}
**Answer:**
{previous_answer}

**Rejection Reason:** {validation_feedback}
{
  • "Too long ({word_count} words)" -> word limit exceeded
  • "Does not cite all essential sentences: {missing_citations}" -> missed citations
  • "Citations to non-essential sentences: {invalid_citations}" -> invalid citations
}

## Instructions for revision
- Review ALL previous rejection reasons
- Ensure ALL essential sentences are properly cited
- Maintain a concise response (maximum {words_limit} words)
- Make sure to address all issues from previous attempts
```

B.1 End-to-End approach prompt

```
# Medical question answering based on clinical notes
## Task
Generate an answer to a patient's health question using only information from the clinical note. Each statement in your answer must be grounded in specific sentences from the note.

1. Generate an answer to the patient's question.
2. Include information that explain medical reasoning, procedures, relevant medical history of the patient that provides a full answer to the question.
3. EVERY sentence in your answer MUST end with at least one citation in parentheses, e.g., "The procedure was performed to treat the condition (3)." or "The treatment involved multiple steps to address your condition (3, 5)."
4. Be precise with your citations - only cite sentences that support each claim.
5. Be accurate with your citations, make sure citation format is correct: (sentence_number) OR (sentence_number_1, sentence_number_2, ...)
    - Invalid citation examples to avoid: (1-3); (1-2, 5-6); (Sentence 2)
    - Valid citation examples instead: (1, 3); (1, 2, 5, 6); (2)
6. Cite at most a couple of sentences at a time, not more.
7. Keep your answer under {words_limit} words total.
8. Do not include any sentences without citations.

[Example showing format with citations...]

## Current Case
**Patient narrative:** {patient_narrative}

**Clinician question:** {clinician_question}

### Clinical Note (numbered sentences):
{numbered_note}

## [For Iteration i > 1] Previous Attempts
### Attempt {i-1}
**Answer:** {previous_answer}
**Rejection Reason:** {validation_feedback}
{
  • "Too long ({word_count} words)" -> word limit exceeded
  • "Sentence {n} doesn't end with citation" -> missing citation
  • "Poorly formatted citation" -> citation format error
  • "Invalid citation numbers" -> cited non-existent sentences
}

Review ALL previous rejection reasons, and do not repeat these mistakes
```