CUNI-a at ArchEHR-QA 2025: Do we need Giant LLMs for Clinical QA?

Vojtěch Lanz and Pavel Pecina
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{lanz,pecina}@ufal.mff.cuni.cz

Abstract

In this paper, we present our submission to the ArchEHR-QA 2025 shared task, which focuses on answering patient questions based on excerpts from electronic health record (EHR) discharge summaries. Our approach identifies essential sentences relevant to a patient's question using a combination of few-shot inference with the Med42-8B model, cosine similarity over clinical term embeddings, and the MedCPT cross-encoder relevance model. Then, concise answers are generated on the basis of these selected sentences. Despite not relying on large language models (LLMs) with tens of billions of parameters, our method achieves competitive results, demonstrating the potential of resource-efficient solutions for clinical NLP applications.

1 Introduction

Responding to patient messages through EHR portals is increasingly recognized as a burden for clinicians (Budd, 2023). To alleviate this problem, the BioNLP 2025 ArchEHR-QA shared task (Soni and Demner-Fushman, 2025b) challenges participants to automatically answer patients' questions using the content of their EHRs. The task requires identifying essential information from an excerpt of a clinical discharge summary and using it to generate accurate and relevant answers.

One of the main limitations of the shared task is the absence of training data, reflecting real-world deployment settings where hospitals often lack the resources to curate and annotate large datasets. Instead, participants are given a small development set consisting of 20 cases. Each case includes a patient's question, a clinician-paraphrased version of the question, and an excerpt of the discharge summary segmented into sentences. The ground-truth annotations identify the sentences essential for answering the question. The final test set comprises 100 similar cases, but without access to ground truth annotations. For a detailed description of the dataset, see Soni and Demner-Fushman (2025a).

Furthermore, healthcare institutions are limited in using external services due to privacy restrictions and, at the same time, cannot easily integrate large-scale LLMs with tens of billions of parameters on-premise due to hardware requirements (Jiang et al., 2023). Therefore, in this submission, we explore approaches that avoid reliance on massive LLMs, focusing instead on lightweight and interpretable components.

Our method combines multiple signal sources to detect essential sentences relevant to the patient's question, including few-shot inference with the Med42-8B model (Christophe et al., 2024), cosine similarity over clinical term representations, and cross-encoder models trained on clinical pair relevance. The selected sentences are then used to generate a concise answer. [1]

2 Related Work

Clinical NLP research has been supported by several large collections of clinical and biomedical texts, such as MIMIC (Johnson et al., 2023) and PubMed (Canese and Weis, 2013). Not only do these datasets serve as the foundation for various shared tasks aimed at extracting relevant information for specific cases or questions, such as the BioASQ Challenge (Tsatsaronis et al., 2015), the TREC Clinical Trials Track 2022 (Roberts et al., 2022), or the ArchEHR-QA 2025, the task we investigate in this paper. In addition to many approaches to biomedical information retrieval, one notable example is the MedCPT model (Jin et al., 2023), which compares embedding representations of abstract articles with those of input queries.

Other notable clinical datasets include n2c2 (Henry et al., 2019), from which the emrQA Ques-

[1] Source code available at https://github.com/lanzv/CUNI-a-at-ArchEHR-QA-2025

BioNLP 2025 Shared Tasks, pages 27–40
August 1, 2025 ©2025 Association for Computational Linguistics

tion Answering dataset (Pampari et al., 2018) is derived. This dataset was used by Lanz and Pecina (2024) to study paragraph retrieval using models such as ClinicalBERT (Alsentzer et al., 2019) and BioBERT (Lee et al., 2019), both of which are pre-trained on English clinical and biomedical text.

In addition, several clinically pre-trained decoder-based language models were introduced to address a wide range of clinical tasks, including BioMistral (Labrak et al., 2024), Med42 (Christophe et al., 2024), or Meditron 3 (Sallinen et al., 2025). However, recent findings (Dada et al., 2025; Lanz and Pecina, 2025) suggest that clinical pretraining is not always essential and that multilingual or general-domain pretraining may be equally or even more beneficial for certain clinical tasks.

3 Methodology

The methodology follows the structure of the shared task, which has two steps: essential sentence retrieval followed by answer generation.

- **Essential Sentence Retrieval:** In this stage, we iterate over all annotated sentences in the clinical documents and compare them with the clinical question (formulated by clinicians, not patients) to retrieve sentences that are essential to answer the question. This step is evaluated using the **Factuality** score, defined as the micro-averaged F1 score of correctly predicted essential sentences.

- **Answer Generation:** Based on prior predictions, we concatenate the retrieved essential sentences into a compact answer, limited to 75 words - an empirically optimal length (Lin et al., 2003; Jeon et al., 2006) and the evaluation cut-off point for the shared task. This stage is scored using the mean of automatic similarity metrics comparing the generated compact answer with gold essential sentences: BLEU (Papineni et al., 2002), ROUGELsum (Lin, 2004), SARI (Xu et al., 2016), BERTScore (Zhang et al., 2020), AlignScore (Zha et al., 2023), and MEDCON (wai Yim et al., 2023) - collectively referred to as the **Relevance** score.

The final overall evaluation measure averages the Factuality and Relevance scores.

3.1 Essential Sentence Retrieval

We explore several approaches that model different aspects of sentence essentiality for clinical question answering. Each method aims to determine whether a given sentence contains essential information to answer a question formulated by a clinician.

Max Cosine Similarity. This method assumes that if a sentence contains terms similar to those in the question, it is more likely to be essential. However, to avoid the influence of stop words and general-domain terms, we focus exclusively on clinical terminology.

First, we use a SciSpaCy model *en_core_sci_sm* (Neumann et al., 2019) to extract clinical terms from both the sentence and the question. Then, for each pair of clinical terms (one from the sentence, one from the question), we compute the cosine similarity of their embeddings using ClinicalBERT. The maximum cosine similarity among all such pairs is taken as the sentence's relevance score.

We apply a threshold to retrieve sentences that are then considered essential. Following Lanz and Pecina (2025), we also test mBERT instead of ClinicalBERT to compare domain-specific and multilingual pretraining. We refer to the resulting methods as *MCS-C* and *MCS-M*, based on ClinicalBERT and mBERT, respectively.

MedCPT Cross-Encoder. Lexical similarity may not capture semantic relevance when different terms convey similar meanings. To address this, we use the MedCPT Cross-Encoder, trained for biomedical information retrieval on PubMed. It takes a sentence–question pair as input and outputs a similarity score, which we threshold to determine the essentiality. We refer to this approach as *MedCPT FS* (Full Sentences).

To reduce noise, we also experiment with filtering non-clinical content using the SciSpaCy extraction model. Both sentences and questions are reduced to comma separated clinical terms before being inputted into MedCPT. This variant is denoted as *MedCPT CT* (Clinical Terms).

Sentence Relevance with Med42-8B. Due to clinical privacy constraints, externally hosted models such as ChatGPT (OpenAI, 2025) cannot be used with MIMIC data – a common limitation in clinical NLP. This requires a secure, local deployment, which is often infeasible in hospitals due to limited infrastructure. As deploying large models is impractical in such settings, we focus on smaller and more efficient models suitable for local use.

Furthermore, the lack of training data implies the use of zero- or few-shot methods. Therefore, we use Med42-8B, a compact, instruction-tuned model that has undergone preference optimization for interactive tasks. Our few-shot prompt includes synthetic examples generated by GPT-4o - each with a patient question, candidate sentence, answer (or None), and justification. Importantly, we ensure that no data from the shared task are included in the few-shot generation process. Otherwise, we could not use the dev set for a fair validation-based comparison of approaches before evaluating the best approach on the final test set. And while it might seem appealing to use real data - or at least data closely resembling it, such as using some of dev set examples as few-shot prompts - this would not only be methodologically incorrect, but also impractical: the dev set is already so small that we must preserve it entirely for validation purposes. Furthermore, we cannot share shared task data with third-party services. Therefore, we rely on synthetic examples generated by GPT-4o shown in Appendix D.

The confidence score for each prediction is computed from the token-level softmax probabilities of the model's output, covering both the answer and its justification. If the model generates None as the answer, the confidence is set to 0.0. The scores obtained within each patient case are normalized by dividing by their total sum; If the sum is 0.0 (that is, all values are zero), no normalization is applied. We refer to this model as *SR Med42*.

Context-aware Relevance with Med42-8B. Previous approaches assessed sentences in isolation, but clinical text often relies on earlier context for full meaning. For example, a sentence *"In that case, notify the cardiology team."* is only relevant to the question *"What should be done if the patient develops chest pain?"*, if we know *"that case"* refers to chest pain, illustrating the need for context-aware relevance.

To incorporate this, we propose *CAR Med42*, which applies Med42-8B with the full summary of discharge. Few-shot prompts, generated via GPT-4o, include a clinical context, patient question, candidate sentence from the context, binary answer (Yes/No), and justification.

As before, a No prediction yields a score of 0.0, while a Yes prediction uses the model's generation probability as the relevance score (and again, if possible, scaling is applied). Importantly, no shared task or clinical data was shared with ChatGPT - only synthetic examples were used. A complete few-shot example is provided in Appendix E.

3.2 Answer Generation

Once the essential sentences are retrieved, they are used to construct the final answer. The goal is to provide a direct response while ensuring that the answer stays under the 75-word limit.

First, each essential sentence is compressed individually. We prompt the Med42-8B model in a few-shot setting (with examples generated by GPT-4o) to generate a concise direct answer using the essential sentence as context. If the model cannot generate an answer, the sentence is shortened with a second few-shot prompt, also with Med42-8B, focusing on compressing the sentence while preserving its content. The corresponding prompt templates are shown in Appendix F and Appendix G, respectively.

After processing all the essential sentences, we concatenate them into a single answer. If the result exceeds the 75-word limit, we iteratively shorten the longest processed sentences using the second few-shot prompt until the word count is within bounds. In rare cases where this process stalls (i.e., no length reduction after two iterations), we remove the last word from the longest sentence and attempt compression again.

Conversely, if the final answer is significantly shorter than the limit, we gradually replace the most concise processed sentences with their original, longer, essential sentence forms. This ensures that the answer contains as much relevant information as possible while remaining easy for patients to understand.

4 Results

The sentence retrieval methods we proposed return confidence scores rather than binary decisions. Although SR Med42 and CAR Med42 explicitly assign a confidence score of 0.0 when the Med42 model predicts that a sentence is not essential, we still need to apply a threshold to convert the scores into final binary decisions. Thus, we first optimize threshold values on the development set and then apply the optimal thresholds to the test set for evaluation.

Although tuning thresholds on the development set of only 20 cases may raise concerns about over-

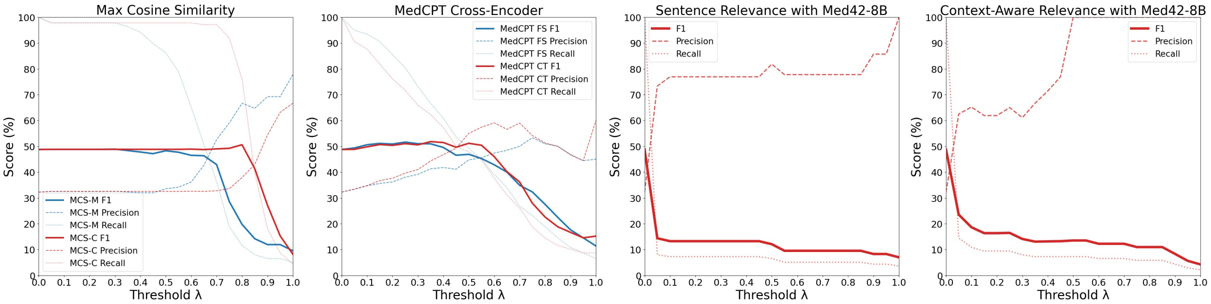

Figure 1: Micro F1, precision, and recall scores across varying confidence thresholds λ for four essential sentence retrieval methods. Only sentences with a model confidence score greater than or equal to λ are considered essential.

fitting, each case contains multiple sentences, resulting in hundreds of sentence-level evaluations. This yields a sufficiently informative signal to guide threshold selection, even if it may not guarantee a globally optimal setting. Crucially, since the threshold is fixed before any test data are seen, the validity of the final test evaluation remains unaffected.

4.1 Threshold Optimization on Dev Set

To identify optimal threshold values for each method, we perform a sweep over a range of threshold values and analyze the resulting precision-recall trade-offs. As shown in Figure 1, higher thresholds improve precision but reduce recall, limiting F1 performance. In fact, F1 scores often do not significantly exceed the baseline for retrieving all sentences as essential.

In the Max Cosine Similarity results, we observe that, while MCS-C achieves higher F1, MCS-M obtains better precision. Similarly, in the MedCPT Cross-Encoder results, both MedCPT FS and MedCPT CT follow similar trends, with the clinical-term-filtered variant (MedCPT CT) performing slightly better. Based on this, we prioritize MedCPT CT in subsequent experiments, as filtering non-clinical content helps reduce noise. However, for Max Cosine Similarity, neither model clearly dominates.

Given the findings that clinical pretraining does not always help (Dada et al., 2025; Lanz and Pecina, 2025), in the SR Med42 and CAR Med42 approaches, we experimented with replacing the Med42-8B model with its base non-medical alternative, Llama3-8B (Grattafiori et al., 2024). However, despite similar trends in the precision, recall, and F1 curves, the general-domain Llama3-8B lags behind Med42-8B (see Figure 4). Therefore, we rely on the Med42-8B model in these approaches.

Method	***Overall***	***Factuality***	***Relevance***
Ensemble-C	48.6	56.8	**40.5**
Ensemble-M	**49.0**	**58.6**	39.4

Table 1: Factuality F1 (***Fact***), Relevance (***Rel***) metrics, and their mean **Overall** score of the two approaches, Ensemble-M and Ensemble-C, measured on the dev set.

Method	**F1**	***Precision***	***Recall***
All Sentences	48.8	32.2	100.0
MCS-C	50.6	37.9	76.1
MCS-M	48.9	32.6	97.8
MedCPT FS	51.6	38.0	80.4
MedCPT CT	51.8	44.3	62.3
SR Med42	48.8	32.2	100.0
CAR Med42	48.8	32.2	100.0
Ensemble-C	56.8	53.2	60.9
Ensemble-M	**58.6**	52.3	66.7

Table 2: Comparison of F1, Precision, and Recall across methods for essential sentence retrieval.

Since each method captures different aspects of sentence essentiality, we explore combining them in ensemble models. A sentence is retrieved as essential if at least one of the selected methods assigns it a score above its respective threshold.

We define two ensembles:

- **Ensemble-C**: combines MCS-C, MedCPT CT, SR Med42, and CAR Med42
- **Ensemble-M**: combines MCS-M, MedCPT CT, SR Med42, and CAR Med42

We then perform a grid search for combinations of thresholds to maximize F1 in the development set (see Appendix A).

Table 2 summarizes the best F1 scores achieved by each method, including the baseline where all sentences are considered essential. The ensemble

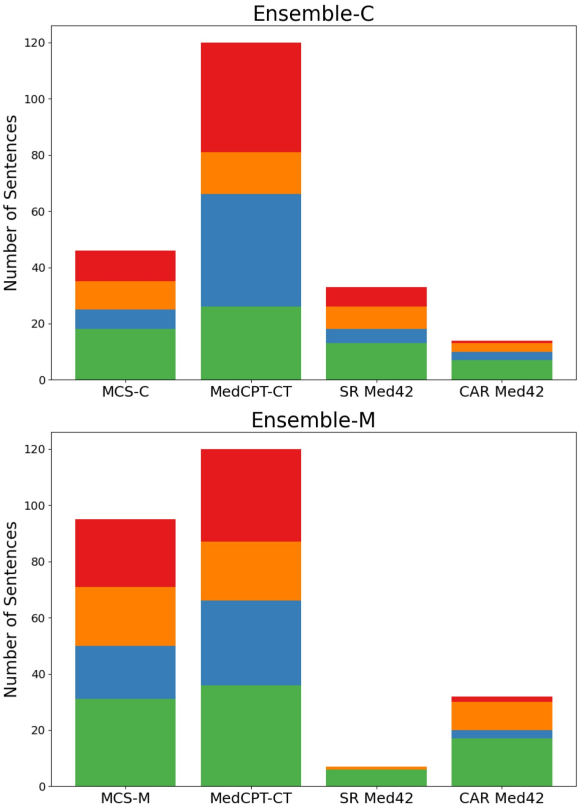

Figure 2: Contribution of individual models within Ensemble-C (top) and Ensemble-M (bottom), showing for each method the number of predicted essential sentences. Bars decompose into: red (unique wrong predictions), orange (wrong and also predicted as wrong by at least one other submethod), blue (correct and unique contribution), and green (correct, but also predicted correctly by at least one other submethod).

methods clearly outperform the individual models. To assess the robustness of these results, we estimate the variability of the F1 scores using bootstrap resampling over the input examples. This involves repeatedly sampling subsets of the data with replacement and re-computing the F1 score on each sample. The resulting distributions yield estimated means and standard deviations of 56.23 ± 3.91 for Ensemble-C and 58.64 ± 3.29 for Ensemble-M, indicating that both ensembles consistently outperform the baselines in the resampled data.

Figure 2 visualizes the contribution of each method within the ensemble approaches to the final prediction of essential sentences. The figure shows that MedCPT-CT is the most dominant contributor. Interestingly, MCS-M plays a much more significant role in Ensemble-M than MCS-C does in Ensemble-C. However, the ratio of correctly and incorrectly predicted sentences remains similar across all methods.

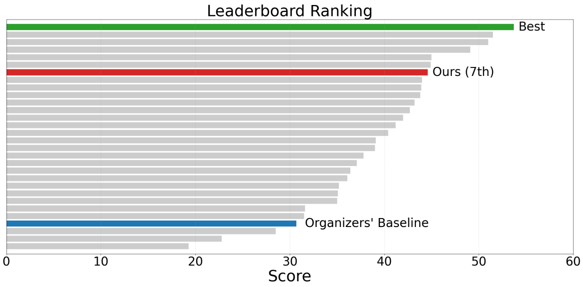

Figure 3: Automatic evaluation on the test set. Our method ranks 7th out of 30 submissions.

The final test set results for the ensemble methods are shown in Table 1. Additional Relevance subscores are reported in Table 3 in Appendix B. While Ensemble-M achieves the highest Factuality score, Ensemble-C performs slightly better in Relevance. Overall, Ensemble-M is the stronger method, and we select it as our final approach for evaluation on the test set.

4.2 Final Test Set Results

Our final system achieved a score of 44.6 on the test set, placing us 7th out of 30 participating teams (see Figure 3). These automatically evaluated results show that our ensemble-based approach is competitive, despite not using LLMs with ten billion or more parameters.

5 Conclusion

In this work, we presented our submission to the ArchEHR-QA 2025 shared task. We focused on identifying essential sentences for answering a given patient's question. Based on these predicted sentences, we generated the final compact answer. We combined a few-shot Med42-8B model with cosine similarities of clinical terms and the MedCPT cross-encoder scores.

Our results are reasonable and competitive, even without using LLMs with tens of billions of parameters, which are not easily integrable into hospital environments. Furthermore, although replacing the domain-specific Med42-8B model with the general-domain Llama3-8B led to a slight drop in performance, it still suggests that domain-specific pretraining provides a modest benefit. However, in the cosine similarity approach, mBERT performs similarly to ClinicalBERT. This highlights that general-purpose multilingual models can still be competitive in clinical tasks.

Limitations

No training data and a small validation set limit the development of the model. The notion of an "essential sentence" is loosely defined and open to interpretation. Our study is limited to English, and few-shot prompts were generated using ChatGPT, which may introduce bias and produce examples that are not fully accurate or tailored to our task. Finally, automatic evaluation may not fully reflect the correctness and clinical validity of the answer.

Acknowledgments

This research was partially supported by the SVV project number 260 698 and the Charles University GAUK grant No. 284125. It has also received support and funding from the European Union's Horizon Europe research and innovation programme project *RES-Q plus* (Grant Agreement No. 101057603). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

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.

Jeffrey Budd. 2023. Burnout related to electronic health record use in primary care. *Journal of Primary Care & Community Health*, 14.

Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1).

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

Amin Dada, Osman Koras, Marie Bauer, Amanda Butler, Kaleb Smith, Jens Kleesiek, and Julian Friedrich. 2025. MeDiSumQA: Patient-oriented question-answer generation from discharge letters. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 124–136, Albuquerque, New Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2019. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, and Soyeon Park. 2006. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 228–235, New York, NY, USA. Association for Computing Machinery.

Xinrui Jiang, Lixiang Yan, Raja Vavekanand, and Mengxuan Hu. 2023. Large language models in healthcare current development and future directions. In *Generative AI Research*.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Vojtech Lanz and Pavel Pecina. 2024. Paragraph retrieval for enhanced question answering in clinical documents. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 580–590, Bangkok, Thailand. Association for Computational Linguistics.

Vojtech Lanz and Pavel Pecina. 2025. When multilingual models compete with monolingual domain-specific models in clinical question answering. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 69–82, Albuquerque, New Mexico. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

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.

Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, and Boris Katz. 2003. What makes a good answer? the role of context in question answering.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

OpenAI. 2025. Chatgpt (april 2025 version). `https://chat.openai.com`. Large language model accessed in April 2025 via https://chat.openai.com.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

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.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, Steven Bedrick, and William R. Hersh. 2022. Overview of the trec 2022 clinical trials track. In *Text Retrieval Conference*.

Alexandre Sallinen, Antoni-Joan Solergibert, Michael Zhang, Guillaume Boyé, Maud Dupont-Roc, Xavier Theimer-Lienhard, Etienne Boisson, Bastien Bernath, Hichem Hadhri, Antoine Tran, Tahseen Rabbani, Trevor Brokowski, Meditron Medical Doctor Working Group, Tim G. J. Rudner, and Mary-Anne Hartley. 2025. Llama-3-meditron: An open-weight suite of medical LLMs based on llama-3.1. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*.

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.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael Alvers, Dirk Weißenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

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.

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.

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 *International Conference on Learning Representations*.

A Technical Details

For the Ensemble-M configuration, a sentence was predicted as essential if it had a confidence score exceeding 0.7, 0.5, 0.9, or 0.05 in at least one of the following approaches: MCS-M, MedCPT FS, SR Med42, and CAR Med42, respectively.

Similarly, in the Ensemble-C setup, a sentence was predicted as essential if it exceeded the thresholds of 0.9, 0.5, 0.0, or 0.4 in at least one of the confidence scores from MCS-C, MedCPT FS, SR Med42, and CAR Med42, respectively.

B Relevance Scores on Dev Set

Method	BLEU	ROUGELsum	SARI	BERTScore	AlignScore	MEDCON
Ensemble-M	7.1	29.5	66.9	34.9	55.0	42.9
Ensemble-C	7.5	31.0	66.3	36.5	59.6	42.0

Table 3: All relevance scores of Ensemble-M and Ensemble-C approaches measured on the dev set.

C Performance Comparison: Med42-8B vs. Llama3-8B

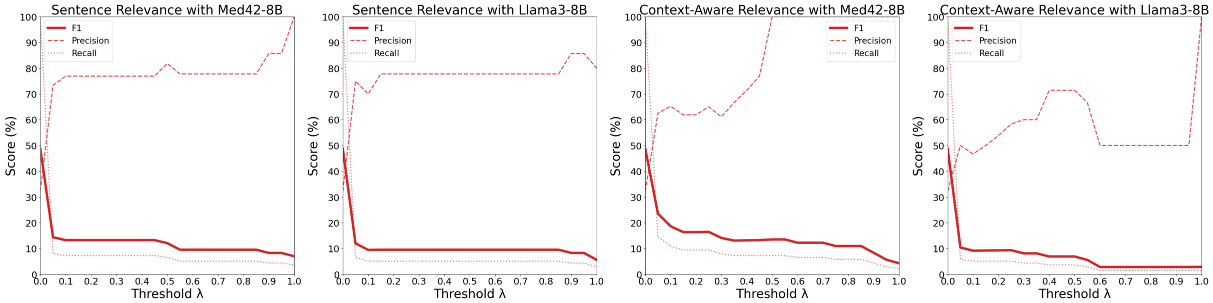

Figure 4: Comparison of sentence retrieval and context-aware relevance performance using Med42-8B vs. Llama-8B. Each chart shows results for one task-model pair, highlighting the impact of replacing Med42-8B with a general-domain Llama-8B model.

D Essential Sentence Retrieval – Few-Shot Prompt for Sentence Relevance Med42

```
You are a clinical assistant. Given a context and a question, extract only the
    essential information from the context that is necessary to answer the question.
     If no information is relevant, respond with "None". Also provide a short
    explanation for your answer.

Context: The patient has a history of hypertension and presents with progressive
    shortness of breath. BNP levels are elevated. Physical examination reveals
    bilateral rales and mild pedal edema.
Question: What information is essential from this context for answering the question
     "What is causing the patient's breathing difficulty?"
Answer: elevated BNP, bilateral rales, mild pedal edema
Reason: These are all indicators of congestive heart failure, which likely explains
    the breathing difficulty.

Context: The patient is a 45-year-old male with a history of allergic rhinitis. He
    was seen in allergy clinic and placed on a regimen of nasal corticosteroids and
    antihistamines. No new triggers identified. Symptoms are seasonal and well-
    controlled.
Question: What information is essential from this context for answering the question
     "What is the most likely cause of the patient's anemia?"
Answer: None
Reason: The context is entirely focused on allergic rhinitis, with no hematologic
    data or symptoms of anemia.

Context: The patient completed a dental cleaning and X-rays showed mild periodontal
    disease. Oral hygiene habits were discussed, and the patient agreed to floss
    daily. No pain or bleeding reported. No antibiotics were prescribed.
Question: What information is essential from this context for answering the question
     "What medications are responsible for the patient's elevated INR?"
Answer: None
Reason: There is no mention of any anticoagulants or medications that affect
    coagulation in the context.

Context: Patient underwent knee replacement two years ago. Reports occasional
    clicking sensation but no pain. X-ray shows proper implant positioning.
```

```
Question: What information is essential from this context for answering the question
     ”Is the knee replacement causing complications?”
Answer: occasional clicking sensation, no pain, proper implant positioning
Reason: Clicking may suggest minor mechanical noise but no signs of complications
    given the lack of pain and good positioning.

Context: Complains of weight loss and fatigue over the past 3 months. Labs show iron
     deficiency anemia. Colonoscopy reveals a 2 cm mass in the ascending colon.
Question: What information is essential from this context for answering the question
     ”What might explain the patient’s fatigue?”
Answer: iron deficiency anemia, 2 cm mass in ascending colon
Reason: Chronic blood loss from the mass could explain anemia and fatigue.

Context: The patient underwent cataract surgery on the right eye and reports
    improved vision. Post-op evaluation showed clear lens placement and normal
    intraocular pressure. No inflammation noted. Scheduled for left eye surgery in
    two months.
Question: What information is essential from this context for answering the question
     ”Why did the patient develop shortness of breath?”
Answer: None
Reason: The context is limited to ophthalmologic findings and does not mention any
    pulmonary or cardiovascular symptoms.

Context: Denies smoking, alcohol, or drug use. Family history positive for lung
    cancer in both parents. Works in construction for 25 years without respiratory
    protection.
Question: What information is essential from this context for answering the question
     ”What are the patient’s risk factors for lung cancer?”
Answer: family history of lung cancer, 25 years in construction without respiratory
    protection
Reason: Occupational exposure and genetics increase risk even without smoking.

Context: Admitted for severe epigastric pain. Has history of NSAID use for chronic
    back pain. Labs show decreased hemoglobin. Endoscopy confirms a gastric ulcer.
Question: What information is essential from this context for answering the question
     ”What is the likely cause of the gastrointestinal bleeding?”
Answer: NSAID use, gastric ulcer, decreased hemoglobin
Reason: NSAIDs are known to cause gastric ulcers, which can lead to bleeding.

Context: No prior psychiatric history. The patient has been irritable and withdrawn
    for the past month. Sleep has decreased to 3 hours/night. Appetite remains
    normal.
Question: What information is essential from this context for answering the question
     ”Are there signs of depression?”
Answer: irritability, social withdrawal, decreased sleep
Reason: These are common symptoms associated with depressive disorders.

Context: The patient had a colonoscopy last week, which revealed three polyps that
    were removed. Pathology is pending. The patient denies abdominal pain, nausea,
    or changes in bowel habits. Family history is negative for colorectal cancer.
Question: What information is essential from this context for answering the question
     ”Why is the patient experiencing chronic fatigue?”
Answer: None
Reason: The context is focused on GI screening and doesn’t include symptoms, labs,
    or findings that would explain fatigue.

Context: Presents with left arm weakness and facial droop for 45 minutes. Symptoms
    resolved prior to arrival. CT scan shows no acute infarct. History of atrial
    fibrillation.
Question: What information is essential from this context for answering the question
     ”What might have caused the neurological symptoms?”
Answer: transient symptoms, atrial fibrillation
Reason: AFib can cause transient ischemic attacks, which present with stroke-like
    symptoms that resolve.

Context: Mother reports that her child, aged 3, has not yet started speaking in full
     sentences. Hearing test is normal. No social interaction issues observed.
    Growth chart is appropriate.
Question: What information is essential from this context for answering the question
     ”Is there concern for developmental delay?”
```

```
Answer: 3-year-old not speaking in full sentences
Reason: While social and hearing are normal, speech delay is suggestive of possible
    developmental delay.

Context: Recent travel to sub-Saharan Africa. Developed intermittent fever and
    chills on return. Blood smear reveals Plasmodium falciparum.
Question: What information is essential from this context for answering the question
     ”What is the likely cause of the patient’s fever?”
Answer: travel to sub-Saharan Africa, Plasmodium falciparum
Reason: These findings point to malaria as the likely cause of the fever.

Context: Complains of morning stiffness lasting more than 1 hour. Joints in both
    hands are swollen and tender. Positive rheumatoid factor and anti-CCP antibodies
    .
Question: What information is essential from this context for answering the question
     ”Is this likely to be rheumatoid arthritis?”
Answer: morning stiffness >1 hour, swollen/tender hand joints, positive RF and anti-
    CCP
Reason: These clinical and serological findings are diagnostic of RA.

Context: A 65-year-old woman was referred to audiology due to recent hearing
    difficulties. Audiogram showed moderate bilateral sensorineural hearing loss.
    Hearing aids were recommended. No signs of vertigo or tinnitus were reported.
Question: What information is essential from this context for answering the question
     ”What led to the patient’s episodes of syncope?”
Answer: None
Reason: The context only contains auditory assessment and does not address
    cardiovascular or neurologic causes.

Context: On insulin therapy. Skipped lunch due to meetings. Found diaphoretic and
    confused. Glucose 42 mg/dL.
Question: What information is essential from this context for answering the question
     ”What explains the patient’s confusion?”
Answer: skipped lunch, insulin therapy, glucose 42 mg/dL
Reason: Hypoglycemia is likely due to missed meal with insulin use.

Context: Reports worsening shortness of breath over 2 weeks. Has COPD. Oxygen
    saturation drops to 89% on ambulation. Chest X-ray shows no infiltrates.
Question: What information is essential from this context for answering the question
     ”What is likely contributing to the patient’s shortness of breath?”
Answer: COPD history, desaturation with ambulation
Reason: COPD with exertional desaturation is a common cause of dyspnea in such
    patients.

Context: Diagnosed with hypothyroidism last year. Currently on levothyroxine.
    Complains of fatigue and cold intolerance. TSH 9.2.
Question: What information is essential from this context for answering the question
     ”Why is the patient still symptomatic?”
Answer: hypothyroidism, TSH 9.2
Reason: Elevated TSH indicates under-replacement with levothyroxine.

Context: Denies any chest pain. Takes beta-blocker for hypertension. EKG reveals
    bradycardia (HR 48 bpm). Patient feels fatigued.
Question: What information is essential from this context for answering the question
     ”What could explain the fatigue?”
Answer: beta-blocker use, bradycardia
Reason: Bradycardia from beta-blockers may result in reduced cardiac output and
    fatigue.

Context: The patient was evaluated in the ophthalmology clinic due to complaints of
    blurry vision. Examination showed no signs of diabetic retinopathy. Blood
    pressure was within normal range. There were no neurological deficits noted.
    Follow-up was scheduled in six months.
Question: What information is essential from this context for answering the question
     ”What is the underlying cause of the patient’s persistent headaches?”
Answer: None
Reason: The context only discusses ophthalmological findings and vision-related
    complaints but contains no information about the cause of headaches.
```

```
Context: Patient presented for a follow-up regarding their post-operative shoulder
    surgery. Physical therapy was recommended and patient reports improvement in
    range of motion. There are no signs of infection or complications. Sleep has
    improved as well.
Question: What information is essential from this context for answering the question
     "What factors contributed to the patient's recent weight loss?"
Answer: None
Reason: The context only discusses orthopedic recovery and makes no mention of diet,
     metabolism, or weight.

Context: The patient was brought in for confusion. No focal neurological deficits
    noted. BUN and creatinine significantly elevated. Recently started lisinopril.
Question: What information is essential from this context for answering the question
     "What could explain the altered mental status?"
Answer: elevated BUN/creatinine, started lisinopril
Reason: Acute kidney injury from ACE inhibitors may lead to uremic encephalopathy.

Context: 65-year-old with chronic low back pain. MRI shows mild degenerative disc
    disease. No nerve compression.
Question: What information is essential from this context for answering the question
     "Is surgery indicated?"
Answer: mild degenerative disc disease, no nerve compression
Reason: Conservative treatment is favored as no surgical lesion is present.

Context: During the dermatology consultation, the patient described new-onset skin
    lesions. The rash appeared on the arms and back, non-pruritic and non-painful.
    No signs of infection were noted. Biopsy was scheduled.
Question: What information is essential from this context for answering the question
     "Why has the patient developed elevated liver enzymes?"
Answer: None
Reason: The context centers around dermatological symptoms with no hepatic or
    metabolic findings provided.

Context: History of mechanical heart valve replacement. INR today is 5.2. No active
    bleeding reported.
Question: What information is essential from this context for answering the question
     "What explains the elevated INR?"
Answer: mechanical valve replacement
Reason: Patients require anticoagulation for valves, which can overshoot and elevate
     INR.

Context: Breast mass noted on exam. Mammogram shows suspicious lesion. Biopsy
    confirms ductal carcinoma in situ.
Question: What information is essential from this context for answering the question
     "What is the diagnosis?"
Answer: ductal carcinoma in situ
Reason: Biopsy provides definitive diagnosis.

Context: Patient with ESRD on dialysis. Missed last two sessions. Complains of
    generalized weakness. Potassium level is 6.8.
Question: What information is essential from this context for answering the question
     "What is the likely cause of weakness?"
Answer: missed dialysis sessions, potassium 6.8
Reason: Hyperkalemia and uremia due to missed dialysis likely explain weakness.

Context: {Sentence}
Question: What information is essential from this context for answering the question
     "{Question}"
Answer: ...
Reason: ...
```

E Essential Sentence Retrieval – Few-Shot Prompt for Context-Aware Relevance Med42

```
You are a medical assistant helping a patient's family member understand the
    discharge summary. The family member asks a general question about the patient's
     condition or expected recovery. From the discharge summary, you are evaluating
    whether a specific sentence is essential to help them understand what they truly
     need to know - even if they didn't ask about it directly.
```

```
For each example, decide:
- Is the sentence important for answering the underlying concern in the question? (”
    Yes” or ”No”)
- Briefly explain why or why not.

### Example 1
Context:
The patient was admitted with signs of dehydration and electrolyte imbalance
    following several days of vomiting and diarrhea. Intravenous fluids and
    potassium replacement were administered. He gradually regained strength and
    tolerated oral intake by day 3. There were no signs of infection. Electrolyte
    levels normalized. He was encouraged to maintain oral hydration and avoid NSAIDs
    . Discharge instructions included dietary recommendations. He is to follow up
    with his primary care physician in one week. The patient lives alone and has
    limited mobility. Transportation services were arranged for follow-up.

Patient’s Question: How long will it take for him to fully recover?
Sentence: ”He is to follow up with his primary care physician in one week.”
Answer: Yes
Reason: The scheduled follow-up provides insight into the expected timeline of
    recovery and monitoring, even though the patient didn’t explicitly ask about
    appointments.

### Example 2
Context:
The patient presented with acute asthma exacerbation. She received nebulized
    albuterol and corticosteroids in the emergency department. Oxygen saturation
    improved over 24 hours. There were no signs of pneumonia. She was discharged
    with a prescription for inhaled corticosteroids and a tapering dose of
    prednisone. She was advised to avoid known triggers such as smoke or allergens.
    Patient reported improved breathing at rest but slight shortness of breath
    during activity. No further imaging was ordered. The pulmonologist will review
    her progress in 10 days.

Patient’s Question: Is she okay to go back to work next week?
Sentence: ”The pulmonologist will review her progress in 10 days.”
Answer: Yes
Reason: The timing of the specialist review is crucial for determining readiness to
    return to work, even though the patient didn’t mention the appointment.

### Example 3
Context:
The patient was admitted for routine laparoscopic cholecystectomy. The surgery was
    uncomplicated. Minimal intraoperative bleeding was noted. Postoperative pain was
     managed with oral analgesics. Bowel function resumed within 24 hours. She
    ambulated independently on post-op day 2. The surgical wound was clean and dry.
    Discharge instructions advised avoiding heavy lifting for two weeks. Follow-up
    scheduled with surgery clinic in 14 days. Patient was in good spirits and eager
    to return to normal activities.

Patient’s Question: What should her recovery look like?
Sentence: ”Discharge instructions advised avoiding heavy lifting for two weeks.”
Answer: Yes
Reason: The lifting restriction is an essential part of understanding the expected
    recovery process, even if not directly requested.

### Example 4
Context:
{Discharge summary excerpt}

Patient’s Question: {Question}
Sentence: ”{Sentence}”
Answer: ...
Reason: ...
```

F Answer Generation – Direct Answering Few-Shot Prompt

```
You are a clinical assistant helping family members understand discharge summaries.
Your task is to answer questions based on long clinical sentences, which may include
     irrelevant information.
Always provide a direct, natural answer that is as concise as possible.
Do not repeat or copy any part of the question in your answer.
Do not begin the answer with phrases like ’’Because...’’ or ’’XYZ was recommended
    because...’’.
If no clear answer is possible, reply with: None

Question: What treatment did the patient receive for pneumonia?
Sentence: The patient was diagnosed with pneumonia and treated with intravenous
    antibiotics and oxygen therapy.
Answer: He was treated with antibiotics and oxygen therapy.

Question: Why is the patient taking insulin?
Sentence: Due to a recent diagnosis of type 2 diabetes, the patient was prescribed
    insulin to manage blood sugar levels.
Answer: He was diagnosed with type 2 diabetes.

Question: What caused the patient’s shortness of breath?
Sentence: The patient’s shortness of breath was likely due to fluid accumulation in
    the lungs caused by heart failure.
Answer: He had lung fluid from heart failure.

Question: What mobility assistance does the patient need?
Sentence: After hip surgery, the patient requires a walker and supervision while
    moving.
Answer: He requires a walker and supervision.

Question: Why was a walking cane recommended to the patient?
Sentence: The patient’s vaccination record was updated during the follow-up visit,
    including influenza and tetanus boosters.
Answer: None

Question: What complications occurred during the patient’s hospital stay?
Sentence: The patient experienced atrial fibrillation, transient confusion, and a
    mild allergic reaction to antibiotics during admission.
Answer: He experienced atrial fibrillation, confusion, and an allergic reaction.

Question: {Question}
Sentence: {Sentence}
Answer: ...
```

G Answer Generation – Sentence Compression Few-Shot Prompt

```
You are a clinical assistant specialized in simplifying discharge summaries.
Your task is to take a long clinical sentence and rewrite it as a shorter, natural,
    and concise sentence that preserves the essential clinical information.
Do not copy the entire sentence or use unnecessary detail. Keep it factual, clear,
    and brief.

Sentence: The patient was admitted to the hospital due to a sudden episode of chest
    pain that occurred while he was gardening.
Compressed: Admitted for sudden chest pain during gardening.

Sentence: Following the MRI scan, the patient was found to have a small herniated
    disc at the L4-L5 level.
Compressed: MRI showed a small herniated disc at L4-L5.

Sentence: The patient has a medical history of hypertension, type 2 diabetes, and
    chronic kidney disease stage 3.
Compressed: History includes hypertension, diabetes, and stage 3 kidney disease.

Sentence: She was prescribed albuterol inhaler to be used as needed for episodes of
    shortness of breath.
Compressed: Prescribed albuterol for shortness of breath as needed.

Sentence: During his hospital stay, the patient developed a mild skin rash likely
    due to a reaction to antibiotics.
```

```
Compressed: Developed mild rash from antibiotics.

Sentence: The patient was advised to follow a low-sodium diet and monitor blood
    pressure regularly at home.
Compressed: Advised low-sodium diet and home blood pressure monitoring.

Sentence: He lives alone but receives weekly assistance from his daughter with
    groceries and medication management.
Compressed: Lives alone with weekly help from daughter.

Sentence: The patient’s vaccination record was updated during the follow-up visit,
    including influenza and tetanus boosters.
Compressed: Received flu and tetanus boosters at follow-up.

Sentence: {Sentence}
Compressed: ...
```