CogStack-KCL-UCL at ArchEHR-QA 2025: Investigating Hybrid LLM Approaches for Grounded Clinical Question Answering

Shubham Agarwal[1], Thomas Searle[1,2,3], Kawsar Noor[2], Richard Dobson[1,2,3]

[1]Department of Biostatistics & Health Informatics, King's College London, London, U.K.
[2]Institute of Health Informatics, University College London, London, U.K.
[3]CogStack Limited, London, U.K.

Correspondence: shubham.agarwal@kcl.ac.uk

Abstract

We present our system for the ArchEHR shared task, which focuses on answering clinical and patient-facing questions grounded in real-world EHR data. Our core contribution is a 2-Stage prompting pipeline that separates evidence selection from answer generation while employing in-context learning strategies. Our experimentation leveraged the open-weight Gemma-v3 family of models, with our best submission using the Gemma-12B model securing 5th place overall on the unseen test set. Through systematic experimentation, we demonstrate the effectiveness of task decomposition in improving both factual accuracy and answer relevance in grounded clinical question answering.

1 Introduction

As the adoption of digital systems in healthcare become ubiquitous, patients will expect to be able pose questions of their recent experiences. Responding to these questions in a rapid, thorough and most importantly safe way will ensure patients are more involved on their care and receive overall improved care.

Effective communication between patients and their healthcare providers is a cornerstone of quality care as it plays a critical role in treatment adherence, recovery, and overall health outcomes (Zolnierek and DiMatteo, 2009). Patient portals have emerged as a key tool for facilitating this communication, providing individuals with direct access to their health information and enabling ongoing interaction with their care teams (Irizarry et al., 2015). Modern patient portals go beyond simple data access—they support secure messaging, prescription refill requests, and delivery of tailored educational materials (Lyles et al., 2020).

A growing body of research highlights that patient engagement through these digital platforms is associated with improved health literacy, better understanding, increased medication adherence, and greater satisfaction with care (Han et al., 2019; Otte-Trojel et al., 2014; Carini et al., 2021; Dendere et al., 2019). Portals enabling record review and follow-up questions have been shown to foster better self-management and reduce conflict in decision-making (Najafi et al., 2022; Shay and Lafata, 2015).

Beyond empowerment, these digital systems help reduce medical errors, improve communication of complex information, and foster trust between patients and providers (Bell et al., 2017; DesRoches et al., 2020). Integrating robust and responsive question-answering capabilities into patient portals offers a promising direction for advancing truly patient-centered care. The potential of conversational agents to further enhance communication and engagement is increasingly recognized, with recent studies showing early but promising results in clinical contexts (Laranjo et al., 2018).

2 Background

2.1 Retrieval Augmented Generation

A widely adopted framework for building question-answering systems is Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), which uses a Causal Large Language Models (LLMs) to generate answers. In this framework, a retriever first selects relevant passages from a knowledge source, which are then passed as context to the LLM that leverages it to generate answers. In the medical domain, RAG has been applied to tasks such as clinical decision support (Zakka et al., 2024), medical literature retrieval (Tian et al., 2024), and patient education (Xiong et al., 2024), generating patient-friendly explanations of medical conditions and procedures (Yang et al., 2025).

2.2 ArchEHR shared task

Responding to patients' queries on portals offers numerous benefits, as discussed in Section 1, but it

BioNLP 2025 Shared Tasks, pages 126–135
August 1, 2025 ©2025 Association for Computational Linguistics

has become a contributor to clinician burden. Automatically generating accurate, safe, and context-aware responses to patient questions using data from EHRs can help alleviate this pressure.
Although RAG offers a compelling framework for this task, it faces several limitations, especially in the medical domain. Generated responses can be incorrect, particularly when retrieved documents are ambiguous or conflicting, leading to hallucinations (Wong et al., 2025). The complexity of RAG systems can make it challenging to trace the reasoning behind generated answers, which is especially critical in medical contexts (Yang et al., 2025).

The ArchEHR shared task (Soni and Demner-Fushman, 2025a), hosted on PhysioNet (Goldberger et al., 2000), proposes a benchmark specifically designed to evaluate grounded question answering in the clinical domain. The task focuses on answering patient-facing questions using evidence from EHR notes, with a strong emphasis on two core criteria: **factuality**, which measures whether the generated answer is supported by cited evidence, and **relevance**, which assesses how well the response aligns with the patient's original query (Soni and Demner-Fushman, 2025c). This aims at addressing the above discussed limitations and thereby advancing safe and trustworthy patient-centered clinical QA systems.

2.2.1 Dataset

The ArchEHR dataset (Soni and Demner-Fushman, 2025b) is constructed using the MIMIC-III database (Johnson et al., 2016), a large, publicly available resource of de-identified ICU records, ensuring realistic clinical language and complexity. Each instance in the dataset contains a patient-posed question, a clinician-refined rewrite, a set of evidence sentences from real clinical notes, and a gold standard answer.

The task requires systems to generate responses grounded in the provided evidence, with citations to specific supporting sentences. The task does not enforce answers to use either or only one of the patient or clinician focused questions.

3 Methodology

We present two groups of approaches for the task:

- Firstly, we prompt LLMs either in 2-Stages or a combined 1-Stage approach, as described in Section 3.1.1 and 3.1.2 respectively.
- Secondly, we experiment with a *classical* sentence embeddings, fine-tuning classifiers for classification of relevant evidence to inform the generation step.

These approaches are similar to a general RAG process, but importantly our retriever step is constrained to only sentences, and to three distinct classes of informativeness for the generated summary, i.e. essential, supplementary or not-relevant sentence classes.

3.1 Approaches

3.1.1 2-Stage Prompt Approach

This approach consists of two stages, each targeting a specific subtask: **Stage 1** – Sentence Classification and Retrieval and **Stage 2** – Generation. For both the stages, an LLM is prompted to perform the specified subtask. Below is a description of the stages:

- **Stage 1:** Given a query, Stage 1 focuses on identifying the most relevant sentences from the clinical notes. By passing focused context to the generation stage, it improves performance as the generation stage focuses on clinically meaningful evidence, leading to more precise and context-aware responses.

- **Stage 2:** Stage 2 performs generation using the filtered context, allowing the model can leverage all its capability effectively to produce accurate, clinically relevant answers. This ensures that the final response is not only coherent but also grounded in the relevant evidence, minimizing the risk of hallucinations or misinformation.

The 2-Stage prompt is the proposed approach in this work. Its ability to break down the task into manageable stages improves the clarity and performance of each step, resulting in a higher performing pipeline. Further details are discussed in Section 5.

3.1.2 1-Stage Prompt and 2-Stage Fine-tuned Classifier Approach

The 1-Stage Prompt combines sentence classification and answer generation into a single prompt to the LLM, requiring the model to both identify relevant evidence and generate a response at once. This approach simplifies and aims to streamline the process by tackling the task as a single coherent objective. This approach also utilises the prompting techniques mentioned in Section 3.2.

The 2-Stage fine-tuned classifier approach follows the same 2 stage structure as the 2-Stage prompting method, but uses a fine-tuned classifier to perform the sentence classification in-place of an LLM. Specifically, we use Sentence-BERT embeddings (Reimers and Gurevych, 2019) to encode sentences and train a classifier to perform the task using the dev test. This approach allows for greater control over the sentence filtering stage and enables fine-tuning on the task-specific data.

3.2 Few Shot Learning

To guide the model through both stages, we leverage in-context learning via few-shot prompting to ensure consistent and contextually accurate outputs. Carefully designed prompts which include a small number of examples, help the model understand the task, distinguish relevant from irrelevant information, and structure its responses appropriately.

Prompt design was iteratively refined based on empirical performance during sentence classification and answer generation phases. Our approach integrates task-specific examples that included varied clinical scenarios to better guide the model, allowing it to also grasp the clinical nuances of the tasks. Appendix A.2 provides our prompt templates.

3.3 Output Guardrails and Format Enforcement

To ensure consistency and adherence to format requirements across both stages, we implement guardrails across both stages of the pipeline. An output parser validates the model's responses in both stages with the expected format criteria. In cases where the initial output fails to adhere to the required format, we utilize an additional parser that leverages an LLM to reattempt answering and formatting. This lowers the probability of a response being discarded by allowing it to be reformatted correctly.

3.4 Pre-Trained Models

In our pipeline, we use the Gemma family of models (Gemma Team et al., 2025), specifically the instruction tuned models. These models are openly available and based off the closed source Google Gemini models. We selected Gemma models due to their strong performance on instruction-following tasks and their demonstrated reasoning capabilities with more manageable parameter sizes. Notably, the Gemma v3 models outperform their predecessors across multiple reasoning tasks (Gemma Team et al., 2025), making them suitable for complex clinical question answering. Our initial experiments also included Mistral 7B v0.2 instruct model (Jiang et al., 2023).
Experimentation utilized a shared university resource machine with 3 Nvidia A100 GPUs via KCL CREATE (King's College London e-Research team, 2025). We also utilised LLama-cpp and GGML / GGUF quantized models for directly running models on locally available hardware.

We attempted to use the Gemma 27B with initial experiments for 1-Stage prompting but found the model refused to consistently return results on the dev set. We did not continue experimenting with this model and do not report results. Similarly, we attempted to use the Qwen 2.5 7B instruct model (Qwen Team, 2024). We did not report the results for it as the performance was poor for all approaches.

3.5 Evaluation

The ArchEHR task is evaluated through cited evidence classification performance representing *Factuality* and the quality of the generated responses using the cited evidence representing *Relevance*.

Factuality is measured through precision, recall and F1 of prediction of each source sentence representing of one of three classes 'essential', 'supplementary', 'not-relevant'. Scoring is *strict* if only 'essential' labels are included or *lenient* if both 'essential' and 'supplementary' sentences are counted towards final calculations.

Relevance uses a collection of n-gram based automated evaluation metrics BLUE (Papineni et al., 2002), ROUGE (Lin, 2004), SARI (Xu et al., 2016) and model based metrics BERTScore (Zhang et al., 2020), AlignScore (Zha et al., 2023) and MEDCON (wai Yim et al., 2023). Scoring generated text for relevance to a provided question can be subjective, but aggregating a range of scores provides some means to automatically evaluate system performance at scale.

While open-domain metrics can give a broad indication of fluency and semantic similarity, MEDCON directly assesses the preservation of medical relevance, offering a more trustworthy signal in safety-critical clinical question answering.

External knowledge was permitted during this task, and a real system would likely include the integration of external knowledge supplementing existing knowledge within the LLM or model approach. For example, a local, regional or national clinical guideline could be referenced by an LLM during a generation if a question involved why a course of action was taken.

Our approach did not use any external knowledge, or external clinical knowledge base such as UMLS (Bodenreider, 2004) or SNOMED CT (Stearns et al., 2001). This is further mentioned in Section 6.

4 Results

4.1 2-Stage Prompt Approach

Our approach was evaluated on the dev set and test set using the specified metrics, and the results demonstrate promising performance for clinical question-answering tasks. As shown in the results table 1, the Gemma 12B model outperformed the other models across all metrics, achieving an overall score of **47.03**. This suggests that the larger models are better equipped to follow instructions and capture the complex relationships and context within clinical data*. While the larger models consistently outperformed the smaller ones*, the smaller models exhibited a strong ability to handle complex clinical data.

4.2 1-Stage Prompt and 2-Stage Fine-tuned Classifier Approach

As shown in Appendix A.1, both the 1-Stage prompt and 2-Stage fine-tuned classifier approaches underperform relative to the 2-Stage prompt approach, especially for the Gemma 12B model which shows a performance decrease of 34.7% and 47.1% respectively. The 1-Stage Prompt approach lags in both factuality and relevance, except for a slight gain in relevance for the Gemma 4B model.

Similarly, the 2-Stage fine-tuned classifier approach is subpar overall except Factuality for Gemma 4B model. Notably, this approach achieves high precision scores (for strict and lenient), with lenient macro precision score of 86.25.

5 Discussion

The development of our approach for the ArchEHR task evolved through several iterations, each building on previous insights. The 1-Stage prompt approach exposed the limitations of a monolithic design, as the LLM struggled with handling both classification and generation simultaneously. To address this, we introduced a 2-Stage fine-tuned classifier approach, which showed promise and achieved high factuality and precision but was constrained by limited data for effective training. With these insights, we adopted the 2-Stage prompt approach, which retained the advantages of task separation without requiring fine-tuning. This approach outperformed the others, delivering stronger results in both factuality and relevance.

This approach mimics the Chain-of-Thought reasoning process (Wei et al., 2022), whereby breaking down the task into smaller, sequential subtasks encourages more structured reasoning, improves factual alignment, and reduces cognitive load on the model, enabling it to perform each step more reliably and accurately. It also provides a more interpretable pipeline where each stage can be independently evaluated, enhancing overall system transparency.

While the proposed approach achieves strong results, it depends heavily on prompt design and the inherent capabilities of the underlying LLM. We further discuss the limitations and future work in the below sections.

6 Conclusions & Future Work

Our work presents a 2-Stage few-shot prompting approach to grounded clinical QA from real-world EHR data. Leveraging the Gemma-v3-12B model, our best approach secures 5th place overall on the unseen test set, demonstrating a good balance between *factuality*, recognising the correct sentences that should be used in the generated answer, and *relevance* the quality of the generated text from the cited evidence. This systematic task decomposition enhances performance along with providing a more transparent method, crucial for sensitive healthcare contexts.

Our future work involves integration of external world knowledge into system responses, either as 'guardrails' or to directly improve system responses. An example of such world knowledge could be clinical guideline that informed or impacted a course of action, but is not directly referenced in the source EHR notes. Secondly, we aim to explore fine-tuning a Casual Large Language Model on a more expansive and curated dataset for sentence classification. This would enhance

*Except Gemma 27B model as discussed in Section 3.4

Table 1: Pipeline performance for 2-stage prompting approach

	Factuality					**Relevance**			Overall score
Model	*Strict*		*Lenient*		Overall			Overall	
	Macro F1	Micro F1	Macro F1	Micro F1	factuality	BLEU	SARI	relevance	
Dev set performance									
Mistral 7B	41.89	38.65	43.38	42.21	38.65	3.44	57.4	35.3	36.98
Gemma 1B	25.41	23.38	29.91	28.76	23.38	3.2	62.99	32.54	27.96
Gemma 4B	36.4	31.9	38.2	37.1	31.9	4.1	65.5	38.5	35.2
Gemma 12B	**51.35**	**49.81**	**51.59**	**48.92**	**49.82**	**8.99**	**71.84**	**44.2**	**47.03**
Test set performance									
Gemma 12B	51.4	47.5	52.1	47.6	47.5	4.7	70.0	42.6	45.0

the quality and consistency of context filtering, thereby improving downstream answer quality and reducing reliance on prompt-based reasoning by the LLM.

We look to integrate the development and testing of these methods as we actively pursue safe and reliable clinical QA over EHRs.

Limitations

Our work is presented as a solution to the ArchEHR shared task, and provides results on a small development and unseen larger test set. Our best method generalises well to the unseen test demonstrating the suitability of our method to the task.

However, the proposed system is limited in a number of ways. Firstly, the task and proposed system assumes that entire sentences are either wholly relevant or useful to a response, representing a form of *extractive* summarisation, whereas it is likely an optimal response will likely be helped to *abstractively* summarise from across one or more partial sentences to generate a response.

Secondly, the dataset is small and only representative of a single provider USA based ICU. Further work could expand evaluation of such systems across health systems and geographies.

Usage of our proposed system in a 'production' environment will likely require extensive use of hardware resources, namely GPU compute. Due to the sensitivity of patient EHR data, clinical providers will likely require patient QA systems that leverage LLM technology to be secure and isolated from other systems alongside adhering to regulatory standard such as HIPPA or GDPR. In deployment of clinical informatics systems it is especially important to balance availability of hardware with model and system performance.

Acknowledgments

This work was supported by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. SA, TS, RD are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RD is also supported by The National Institute for Health Research University College London Hospitals Biomedical Research Centre. This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Sigall K Bell, Tom Delbanco, Joann G Elmore, Paul S Fitzgerald, Alan Fossa, Katharine Harcourt, and Suzanne G Leveille. 2017. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. *JAMA Network Open*, 320(18):1867–1878.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Elena Carini, Luca Villani, Anna Maria Pezzullo, Antonio Gentili, Antonella Barbara, Walter Ricciardi, and Stefania Boccia. 2021. The impact of digital patient portals on health outcomes, system efficiency, and patient attitudes: Updated systematic review. *Journal of Medical Internet Research*, 23(9):e26189.

Ronald Dendere, Christine Slade, Andrew Burton-Jones, Clair Sullivan, Andrew Staib, and Monika Janda. 2019. Patient portals facilitating engagement with inpatient electronic medical records: A systematic review. *Journal of Medical Internet Research*, 21(4):e12779.

Catherine M DesRoches, Sigall K Bell, Zhaohui Dong, Joann G Elmore, Paul Fitzgerald, Katharine Harcourt, and Suzanne G Leveille. 2020. Patients managing medications and reading their visit notes: A survey of opennotes participants. *Annals of Internal Medicine*, 172(1):35–38.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *arXiv [cs.CL]*.

A L Goldberger, L A Amaral, L Glass, J M Hausdorff, P C Ivanov, R G Mark, J E Mietus, G B Moody, C K Peng, and H E Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–20.

Haixia Han, Kelly Gleason, Ruopeng Sun, Heather Miller, Huong Kang, Yan Gai, and David Rosenthal. 2019. Using patient portals to improve patient outcomes: Systematic review. *JMIR Human Factors*, 6(4):e15038.

Taya Irizarry, Annette DeVito Dabbs, and Christine R Curran. 2015. Patient portals and patient engagement: a state of the science review. *Journal of medical Internet research*, 17(6):e148.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

King's College London e-Research team. 2025. King's computational research, engineering and technology environment (CREATE).

Liliana Laranjo, Adam G Dunn, Helen L Tong, Ahmet Baki Kocaballi, Jing Chen, Rifat Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, and Enrico Coiera. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Courtney R Lyles, Eugene C Nelson, Susan Frampton, Patricia C Dykes, Anupama G Cemballi, and Urmimala Sarkar. 2020. Using electronic health record portals to improve patient engagement: research priorities and best practices. *Annals of internal medicine*, 172(11_Supplement):S123–S129.

Faezeh Najafi, Pirhossein Shojaei, Saeed Shojaee Moghaddam, Mehdi Jafari, and Arash Rashidian. 2022. Impact of patient engagement on healthcare quality: A scoping review. *Annals of Global Health*, 88(1):78.

Tine Otte-Trojel, Antoinette de Bont, Joris van de Klundert, and Thomas G Rundall. 2014. How outcomes are achieved through patient portals: a realist review. *Journal of the American Medical Informatics Association*, 21(4):751–757.

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. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

L Aubree Shay and Jennifer Elston Lafata. 2015. Where is the evidence? a systematic review of shared decision making and patient outcomes. *Medical decision making*, 35(1):114–131.

Sarvesh Soni and Dina Demner-Fushman. 2025a. ArchEHR-QA: BioNLP at ACL 2025 shared task on grounded electronic health record question answering.

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

Sarvesh Soni and Dina Demner-Fushman. 2025c. 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.

M Q Stearns, C Price, K A Spackman, and A Y Wang. 2001. SNOMED clinical terms: overview of the development process and project status. *Proc. AMIA Symp.*, pages 662–666.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, and 1 others. 2024. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Lionel Wong, Ayman Ali, Raymond Xiong, Shannon Zeijang Shen, Yoon Kim, and Monica Agrawal. 2025. Retrieval-augmented systems can be dangerous medical communicators. *arXiv preprint arXiv:2502.14898*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

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.

Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. 2025. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems*, 2(1):2.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, and 1 others. 2024. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068.

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*, pages 11300–11316. 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 (ICLR)*.

Kelly B Haskard Zolnierek and M Robin DiMatteo. 2009. Physician communication and patient adherence to treatment: a meta-analysis. *Medical care*, 47(8):826–834.

A Appendix

A.1 Results for all approaches

Table 2: Pipeline performance for 1-Stage prompt approach

Dev set Performance

Model	Factuality					Relevance			Overall score
	Strict		*Lenient*		Overall factuality	BLEU	SARI	Overall relevance	
	Macro F1	Micro F1	Macro F1	Micro F1					
Gemma 4B	24.66	25.25	24.65	24.89	25.25	3.57	68.63	41.79	33.52
Gemma 12B	21.10	22.56	21.03	21.13	22.56	1.36	65.70	38.84	30.70

Table 3: Pipeline performance for 2-Stage approach with fine-tuned classifier

Dev set Performance

Model	Factuality							Relevance	Overall score
	Strict				Overall factuality	BLEU	SARI	Overall relevance	
	Macro Precision	Macro F1	Micro Precision	Micro F1					
Gemma 4B	72.08	35.0	71.42	37.43	37.43	1.32	56.38	27.932	32.67
Gemma 12B	70.0	22.93	69.56	19.8	19.8	2.99	62.29	29.89	24.88

A.2 Prompt for 2-Stage pipeline

A.2.1 For Stage 1

```
"""
<bos><start_of_turn>user You are a clinical assistant. Use the context below to perform the given
     task. Your response must be a JSON list of citations.

Answer using the given context to help, if you don’t know the answer, just say that you don’t know,
     don’t try to make up an answer.
Format of Context:
ID: <chunk ID> ; text : <text>
ID: <chunk ID> ; text : <text>
...

The context contains all sentences from note excerpts. These sentences have two categories: relevant
     and not relevant.
Your task is to using reasoning and filter out the ones that are relevant to the question, and
     respond with their ID. Ensure to pick all the relevant ones, prioritise higher recall over
     precision.
Include all chunks that are directly relevant and reasonably connected to answering the question.
     Only exclude chunks that are clearly unrelated.

Output format:
Your output must be a list of structured objects with:
    - ’citation’: the chunk ID (e.g., ’1’)

    - ’citation’: the chunk ID (e.g., ’2’)

    - ’citation’: the chunk ID (e.g., ’4’)

DO NOT add explanations, only the above output.
NOTE: The sentences may not be directly relevant, you will have to infer it.

Examples:
```

```
# Example 1:
**Context:**
ID: 1 ; Text: "The patient complained of frequent urination and excessive thirst. Laboratory tests
     revealed elevated blood glucose levels."
ID: 2 ; Text: "The patient was diagnosed with type 2 diabetes mellitus."
ID: 3 ; Text: "Dietary counseling was initiated to help manage blood sugar levels."
ID: 4 ; Text: "The patient also reported occasional headaches over the past month."

**Question:** What is the patient’s diagnosis?
**Answer:**
[{{"citation": "2"}},
{{"citation": "1"}},
{{"citation": "3"}}]
**Reasoning:
ID 2 gives the direct diagnosis (must include).
ID 1 gives symptoms and test results leading to diagnosis (should include).
ID 3 mentions management for blood sugar slightly grey, but include as it supports the context of the
      diagnosis.
ID 4 about headaches is unrelated (exclude).

# Example 2:
**Context:**
ID: 1 ; Text: "The patient sustained a fractured right femur after a fall from a ladder."
ID: 2 ; Text: "An open reduction and internal fixation (ORIF) surgery was performed to stabilize the
     fracture."
ID: 3 ; Text: "The patient was prescribed physical therapy after hospital discharge."
ID: 4 ; Text: "The patient’s blood pressure was also found to be elevated during admission."

**Question:** What treatment did the patient receive for the femur fracture?
**Answer:**
[{{"citation": "2"}},
{{"citation": "3"}},
{{"citation": "1"}}]
**Reasoning:
ID 2 describes surgical treatment (must include).
ID 3 is post-surgical physical therapy (treatment-related; include).
ID 1 gives context about the fracture itself include because it’s important background to understand
     the treatment.
ID 4 about blood pressure is unrelated (exclude).

Context: {context}

Question: {query}

DO NOT add explanations, only the mentioned output <end_of_turn>
<start_of_turn>model
"""
```

A.2.2 Prompt for Stage 2

```
"""
<bos><start_of_turn>user You are a clinical assistant. Use all of the context below to answer the
     question. Your response must be a JSON list of sentence-grounding pairs.
Answer the question using the given context to help, if you don’t know the answer, just say that you
     don’t know, don’t try to make up an answer.

Format of Context:
ID: <chunk ID> ; text : <text>
ID: <chunk ID> ; text : <text>
...

Output format:
Your output must be a list of structured objects with:
    - ’statement’: part of the response
    - ’citation’: the chunk ID (e.g., ’1’) it came from to ground it in evidence

    - ’statement’: part of the response
    - ’citation’: the chunk ID (e.g., ’2’) it came from to ground it in evidence
    - ’statement’: part of the response
    - ’citation’: the chunk ID (e.g., ’4’) it came from to ground it in evidence

DO NOT add explanations, only the above output.
NOTE: Use all of the sources and cite all sources, do not omit any one, all are relevant.

Examples

# Example 1:
**Context:**
ID: 1 ; Text: "The patient complained of frequent urination and excessive thirst. Laboratory tests
     revealed elevated blood glucose levels."
ID: 2 ; Text: "The patient was diagnosed with type 2 diabetes mellitus."

**Question:** What is the patient’s diagnosis?
**Answer:**
[{{"statement": "The patient was diagnosed with type 2 diabetes mellitus.", "citation": "2"}},
{{"statement": "Laboratory tests revealed elevated blood glucose levels.", "citation": "1"}}]

# Example 2:
**Context:**
ID: 1 ; Text: "An open reduction and internal fixation (ORIF) surgery was performed to stabilize the
     fracture."
ID: 2 ; Text: "The patient was prescribed physical therapy after hospital discharge."

**Question:** What treatment did the patient receive for the femur fracture?
**Answer:**
[{{"statement": "The patient underwent open reduction and internal fixation (ORIF) surgery to
     stabilize the femur fracture.", "citation": "1"}},
{{"statement": "The patient was prescribed physical therapy after hospital discharge.", "citation":
     "2"}}]

Context: {context}
Question: {query}

You can combine the sentences too, there is a word limit , so be succinct.
DO NOT add explanations, only the mentioned output.
USE ALL SOURCES, ALL OF THEM ARE IMPORTANT. <end_of_turn>
<start_of_turn>model
"""
```