heiDS at ArchEHR-QA 2025: From Fixed-k to Query-dependent-k for Retrieval Augmented Generation

Ashish Chouhan and **Michael Gertz**
Data Science Group, Institute of Computer Science
Heidelberg University, Germany
{chouhan, gertz}@informatik.uni-heidelberg.de

Abstract

This paper presents the approach of our team called heiDS for the ArchEHR-QA 2025 shared task. A pipeline using a retrieval augmented generation (RAG) framework is designed to generate answers that are attributed to clinical evidence from the electronic health records (EHRs) of patients in response to patient-specific questions. We explored various components of a RAG framework, focusing on ranked list truncation (RLT) retrieval strategies and attribution approaches. Instead of using a fixed top-k RLT retrieval strategy, we employ a query-dependent-k retrieval strategy, including the existing surprise and autocut methods and two new methods proposed in this work, autocut* and elbow. The experimental results show the benefits of our strategy in producing factual and relevant answers when compared to a fixed-k.

1 Introduction

Electronic Health Records (EHRs) are essential in any healthcare system, serving as repositories of the medical history of patients (Häyrinen et al., 2008). Since 2020, patient portals have increased, resulting in more virtual communications between patients and clinicians (Small et al., 2024). As a result, responding to inquiries of patients has become an important issue. Clinicians are reported to spend around 1.5 hours each day managing approximately 150 messages (patient questions) (Small et al., 2024; Liu et al., 2024b). Thus, answering patient-specific questions is a crucial task that relies on information managed in EHRs.

Large Language Models (LLMs) can automate answer generation for patient questions, as these models are trained on extensive textual data (Liu et al., 2024b). However, LLMs are also prone to hallucinations, that is, they may generate answers not supported by a reliable source. This can undermine user trust and potentially harm patients by giving incorrect advice (Huang et al., 2024b). Therefore, attribution, i.e., linking elements of a generated answer to sources, is critical to ensure that every claim is grounded in medical evidence.

Attribution has gained significant attention across various domains, such as the legal and medical domains (Trautmann et al., 2024; Malaviya et al., 2024). Li et al. (2023) outline three approaches for generating answers with attribution. The first approach is direct model-driven attribution, where an LLM generates answers with their sources without using additional information. This is accomplished by fine-tuning or training the model to generate answers that include attributions (Zhang et al., 2024; Patel et al., 2024; Huang et al., 2024a). However, a common issue with this approach is the hallucination of references (Agrawal et al., 2024). The second approach is known as post-retrieval attribution or retrieve-and-read. It retrieves evidence relevant to a query, generating an answer based on that evidence. The LLM is prompted to reference the retrieved information, thereby enforcing attribution (Menick et al., 2022; Nakano et al., 2021; Şahinuç et al., 2024; Gao et al., 2023b). Post-generation attribution (Gao et al., 2023a; Ramu et al., 2024; Cohen-Wang et al., 2024) is the third approach, and it allows the LLM to generate answers without prior attribution and in a post-processing step map answer text back to its sources.

The objective of the BioNLP Grounded Electronic Health Record Question Answering shared task (ArchEHR-QA) (Soni and Demner-Fushman, 2025b) is to generate answers to patient questions, considering clinical note excerpts and attributing them with relevant evidence from the excerpts. Our approach focuses on developing a pipeline for attributed answer generation by employing a retrieval augmented generation (RAG) framework. We experimented with different methods based on the post-retrieval and post-generation attribution ap-

BioNLP 2025 Shared Tasks, pages 50–61

proaches on the ArchEHR-QA development set, which are detailed in Section 2.

2 Pipeline Overview

Our proposed pipeline utilizes a RAG framework to solve the ArchEHR-QA task. This task involves answering health-related questions from patients and providing attributions based on the patients' clinical notes. In this section, we introduce our different methods, including the pipeline we submitted to the ArchEHR-QA 2025 leaderboard. Section 2.1 provides information about the dataset used for the shared task, followed by Section 2.2 describing the baseline. Section 2.3 provides information on our submitted pipeline, which is based on a surprise (Bahri et al., 2023) Ranked List Truncation (RLT) retrieval strategy. Finally, other methods we experimented with (other than the baseline and submitted pipeline) are outlined in Section 2.4.

2.1 Dataset

The dataset for the ArchEHR-QA 2025 shared task, available on PhysioNet[1] (Soni and Demner-Fushman, 2025a), comprises 20 case studies in the development (dev) set and 100 case studies in the test set[2]. Each case study consists of a hand-curated patient question, its corresponding clinician-rewritten version (i.e., clinician question), and excerpts from the patient's clinical notes. See Appendix A for an example of a case study from the dev set and Appendix B for some statistics on the clinical note excerpts. For every sentence in a clinical note, a 1024 dimensional embedding is computed using the `BAAI/bge-large-en-v1.5`[3] model and stored in a FAISS index (Johnson et al., 2019) for semantic search.

2.2 Our Baseline

While we experimented with various retrieval and prompting strategies within the RAG framework, our baseline follows a post-retrieval attribution approach. This involves prompting an LLM to generate answers based on both patient and clinical questions, along with **all** sentences of the clinical note excerpts from the case study. The decisions made for the baseline and other pipelines proposed in this work are supported by experiments that include

- a query that is constructed using both patient and clinical questions instead of considering only one of them (see Appendix D),
- a one-shot prompting approach instead of zero-shot prompting (see Appendix E),
- different LLMs for answer generation with attributions, which are `LLaMA-3.3-70B`[4] and `Mixtral-8x7B`[5] (Dada et al., 2025; Kweon et al., 2024), and
- a maximum number of 200 tokens generated by the LLM (see Appendix F).

On the other hand, the organizers' baseline used the LLaMA-3.3-70B model in a zero-shot prompting approach, where the model is prompted to generate answers that include attributions. If a response is invalid, e.g., exceeding the word limit or lacking valid attribution, the model is again prompted to generate an answer. This is repeated up to five times to obtain a valid output.

2.3 Submitted Pipeline: Surprise Ranked List Truncation (RLT) Retrieval Strategy

The pipeline we submitted for the shared task aligns with baselines utilizing a post-retrieval attribution approach. In this approach, for a query that combines patient and clinical question, semantically similar sentences from the excerpts of clinical notes are retrieved. The similarity score between the query and each sentence is computed using cosine similarity. During retrieval, k represents the number of highest-scoring (top-k) sentences similar to the query. Instead of using a fixed value for k, our team employed a query-dependent-k selection strategy based on the Ranked List Truncation (RLT) method, referred to as "surprise". This method determines the number k of sentences to consider by first adjusting retrieval scores using generalized Pareto distributions from extreme value theory (Pickands, 1975). It truncates a ranked list using a score threshold, allowing for a variable number of relevant sentences to be selected per query (Meng et al., 2024). The selected sentences and query are passed to the LLMs for answer generation, where the model generates an answer with attribution explicitly referencing retrieved sentences from a clinical note.

[1] `https://doi.org/10.13026/zzax-sy62` (accessed on 30th April 2025)

[2] All experiments described in Section 3 use the dev set.

[3] `https://huggingface.co/BAAI/bge-large-en-v1.5` (accessed on 4th May 2025)

[4] `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct` (accessed on 4th May 2025)

[5] `https://huggingface.co/mistralai/Mixtral-8x7B-v0.1` (accessed on 4th May 2025)

Table 1: Retrieval performance on the development set under strict (essential only) and lenient (essential + supplementary) variants. The Strategy and Variant columns list different retrieval strategies and their parameters. Columns P, R, and F1 quantify precision, recall, and F1-score under both variants. The seven best approaches by combined strict and lenient F1-scores (excluding the $k = 54$ row) are highlighted in **bold**.

Strategy	Variant	Strict			Lenient		
		P	R	F1	P	R	F1
fixed-k	$k = 3$	0.53	0.32	0.36	0.70	0.29	0.39
	$\mathbf{k = 10}$	0.43	0.71	**0.50**	0.56	0.71	**0.58**
	$\mathbf{k = 15}$	0.38	0.81	**0.49**	0.51	0.82	**0.60**
	$\mathbf{k = 20}$	0.35	0.89	**0.49**	0.47	0.88	**0.59**
	$k = 54$	0.33	1.00	0.49	0.45	1.00	0.60
fixed-k + re-rankers	**FlashRank** ($k = 20, n = 10$)	0.38	0.68	**0.45**	0.51	0.67	**0.54**
	Cohere ($k = 20, n = 10$)	0.38	0.67	0.45	0.50	0.66	0.53
autocut	—	0.58	0.22	0.27	0.68	0.21	0.28
autocut*	—	0.59	0.35	**0.34**	0.74	0.32	**0.38**
surprise	—	0.36	0.64	**0.42**	0.48	0.62	**0.49**
elbow	—	0.48	0.66	**0.50**	0.62	0.63	**0.55**

2.4 Other Methods

In this section, we outline various methods within the RAG framework by varying its components, namely retrieval strategies and attribution approaches, to assess their impact on performance. We experimented with retrieval strategies other than surprise, including fixed-k, fixed-k and re-ranking, and query-dependent-k strategies like autocut, autocut*, and elbow.

The **Fixed-k** strategy applies a fixed cut-off for all query results, using common values of 3, 10, 15, 20, and 54. **Fixed-k and re-ranking** is a two-step retrieval that first retrieves semantically k similar candidates based on a fixed cut-off. A relevance score is assigned in the second step, selecting top-n (where $n \leq k$) sentences using re-rankers like flashrank (Damodaran, 2023) and cohere[6]. **Autocut**[7] limits candidate sentences based on discontinuities in the computed similarity scores. It determines the first divergence from a straight decline, excluding candidates beyond this point, although it may struggle with uniformly decreasing scores. In this work, we propose **autocut***, a new cut-off strategy that inspects how much each similarity score decreases compared to the previous score, automatically determining cut-offs based on significant changes without any manual adjustments. We also introduce the **elbow** strategy adapted from the elbow method in clustering to determine cut-offs by plotting similarity scores and locating the "elbow" where the transition from high to low relevance occurs, again with no need for preset parameters.

Along with different retrieval strategies, post-generation and post-retrieval attribution approaches have also been tried. In **post-generation attribution**, after a model generates an answer, those retrieved sentences are identified that support each answer sentence by measuring three similarity types: lexical (ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005)), fuzzy (character-based matching), and semantic (BERTScore (Zhang et al., 2020)). Each similarity is assigned a weight w_i, and a combined score is calculated. If this score exceeds a predefined threshold, the candidate sentence is attributed to the generated sentence. This ensures that every claim is explicitly grounded in some original clinical evidence. Detailed setups and results on weight and threshold settings are provided in Appendix G.

The **post-retrieval attribution** approach associates sentence identifiers with each retrieved sentence for attribution during answer generation. Post-processing steps are applied to generated answers to ensure that attributions are properly placed and no irrelevant attributions occur.

3 Experiments

All experiments were conducted on Google Colab[8] using a Tesla T4 GPU (12GB memory)[9]. For ac-

[6]https://docs.cohere.com/docs/rerank-overview (accessed on 4th May 2025)

[7]https://weaviate.io/developers/weaviate/api/graphql/additional-operators#autocut (accessed on 4th May 2025)

[8]https://colab.research.google.com/ (accessed on 4th May 2025)

[9]Code for the proposed pipeline is available online: https://github.com/achouhan93/heiDS-ArchEHR-QA-2025

Table 2: Pipeline evaluation on the development set under one-shot prompting, 200-token limit, and patient+clinician query. Metrics: strict Precision (P), Recall (R), F1-score (F1), overall relevance score in *Relevance* column, and overall pipeline score in *Overall* column. The performance of the organizer baseline, our baseline, top three proposed pipelines, and other experimented pipelines are listed here.

Retrieval	Attribution	Model	P	R	F1	Relevance	Overall
Organizer Baseline		LLaMA-3.3-70B	**0.63**	**0.33**	**0.43**	0.29	**0.36**
Our Baseline		LLaMA-3.3-70B	0.54	0.27	0.36	0.33	0.35
Top Three Proposed Pipelines							
surprise	Post-retrieval	LLaMA-3.3-70B	0.62	0.26	0.37	**0.35**	**0.36**
elbow	Post-retrieval	LLaMA-3.3-70B	0.59	0.27	0.37	0.32	0.35
fixed-k = 15	Post-retrieval	LLaMA-3.3-70B	0.59	0.25	0.35	0.34	0.35
Other Experimented Pipelines							
fixed-k = 10	Post-retrieval	LLaMA-3.3-70B	0.58	0.27	0.37	0.33	0.35
fixed-k = 10	Post-retrieval	Mixtral-8x7B	0.27	0.15	0.19	0.29	0.24
fixed-k = 15	Post-retrieval	Mixtral-8x7B	0.28	0.15	0.19	0.29	0.25
fixed-k = 20	Post-retrieval	LLaMA-3.3-70B	0.51	0.28	0.36	0.35	0.35
fixed-k = 20	Post-retrieval	Mixtral-8x7B	0.30	0.14	0.19	0.28	0.24
fixed-k = 20 + FlashRank	Post-retrieval	LLaMA-3.3-70B	0.52	0.22	0.31	0.34	0.33
fixed-k = 20 + FlashRank	Post-retrieval	Mixtral-8x7B	0.22	0.12	0.15	0.28	0.22
autocut*	Post-retrieval	LLaMA-3.37B	0.57	0.14	0.23	0.32	0.27
autocut*	Post-retrieval	Mixtral-8x7B	0.44	0.12	0.18	0.27	0.23
surprise	Post-retrieval	Mixtral-8x7B	0.33	0.17	0.22	0.29	0.26
elbow	Post-retrieval	Mixtral-8x7B	0.43	0.15	0.22	0.29	0.26
fixed-k = 54	Post-generation	LLaMA-3.3-70B	0.35	0.22	0.27	0.35	0.31

cessing LLMs, we used `InferenceClient`[10] from the `huggingface_hub` library.

3.1 Evaluation Criteria

The development set provided by the organizers includes clinical note excerpts annotated with sentence numbers for attribution. Furthermore, each sentence is labeled as "essential", "supplementary", or "not-relevant". Evaluation is carried out for two variants, a "strict" variant (considering only "essential" labels) and a "lenient" variant (considering both "essential" and "supplementary" labels). Retrieval performance is measured by precision, recall, and F1-score for each variant. The results are shown in Table 1. We selected fixed-k (10, 15, 20), autocut*, surprise, and elbow for downstream answer generation based on these metrics.

We used the official ArchEHR-QA evaluation script for the overall pipeline evaluation to assess *factuality* and *relevance*. *Factuality* is measured by the precision, recall, and F1-score of cited evidence versus ground-truth annotations computed under both variants. *Relevance* compares generated answer sentences to the ground-truth essential sentences using BLEU, ROUGE, SARI (Xu et al., 2016), BERTScore, AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023). The overall relevance score is the average of these metrics, and the final pipeline score is the mean of overall factuality (strict variant F1-score) and overall relevance.

3.2 Comparative Pipeline Evaluation

Building on the ablations in Appendices D–F, we fixed the query (patient + clinician question), one-shot prompt, and 200-token limit, and evaluated our pipeline with two LLMs, `LLaMA-3.3-70B`[11] and `Mixtral-8x7B`[12], under both post-retrieval and post-generation attribution workflows. The results are shown in Table 2.

Post-Retrieval Attribution Evaluation. We paired each of our selected retrieval strategy (fixed-k = 10, 15, 20; autocut*; surprise; elbow) with each LLM and measured strict variant F1-score and overall relevance. Table 2 shows that `LLaMA-3.3-70B` combined with the surprise retrieval strategy achieves a strict F1-score of 0.37 and overall relevance of 0.35, making it our top

[10] `https://huggingface.co/docs/huggingface_hub/v0.30.2/en/package_reference/inference_client` (accessed on 4th May 2025)

[11] `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct` (accessed on 4th May 2025)

[12] `https://huggingface.co/mistralai/Mixtral-8x7B-v0.1` (accessed on 4th May 2025)

post-retrieval configuration, compared to the baselines.

Post-Generation Attribution Evaluation. Using a fixed-k of 54, we varied lexical/fuzzy/semantic weights and threshold values for the LLaMA-3.3-70B model. As shown in Table 5 in Appendix G, the optimal weighting ($w_1 = 0.0, w_2 = 0.5, w_3 = 0.5$, threshold = 0.5) yields a strict F1-score of 0.27 and overall relevance score of 0.35. Although this setup performs best among post-generation configurations, it underperforms relative to the best-performing post-retrieval configuration.

3.3 Pipeline Performance Analysis

While our best-performing pipeline based on the surprise retrieval strategy and post-retrieval attribution achieves a comparable overall score, it does not outperform the organizer's baseline. This outcome can be because of the following factors:

- **Prompt sensitivity of LLMs.** Salinas and Morstatter (2024) demonstrate that even a small perturbation in prompts can cause changes in an LLM's output. Although the organizer baseline and our best-performing pipeline use the same model (LLaMA-3.3-70B), the organizer baseline employs a zero-shot prompt, whereas our pipeline uses a one-shot prompt with stricter formatting and attribution instructions for the model to follow. These subtle prompt design choices could have influenced the model's ability to generate high quality answers with relevant attributions.

- **Difference in context size.** The development set contains up to 54 clinical note excerpt sentences per case study (see Figure 1b), allowing the organizer baseline to input all sentences to LLM as context, thus ensuring a high recall. In contrast, our pipeline relies on a query-dependent-k retrieval method to select a smaller subset of sentences. This approach naturally reduces recall, as some relevant content may not be retrieved, which thus negatively impacts the overall score.

Despite not outperforming the organizer baseline overall score, our pipeline design is motivated by practical considerations for real-world applications. While using all clinical note sentences is feasible within the shared task environment, real-world applications can contain far more text. We consider including complete texts as often infeasible due to LLMs input length constraints and degradation in model performance due to irrelevant information (Shi et al., 2023; Liu et al., 2024a). In such settings, a retrieval step is required, and determining a fixed k that is suitable for all cases is time-consuming. Query-dependent-k retrieval strategies remove the need for manual k tuning by determining the cut-off point based on score distributions. This allows the system to adapt to different types of queries.

4 Conclusion and Discussion

This work explored various RAG framework components for generating answers with attributions to clinical note excerpts. Our research highlights that the best-performing pipeline employs a post-retrieval attribution approach, utilizing the "surprise" RLT strategy and the LLaMA-3.3-70B model. We achieved a strict variant precision of 0.62 and recall of only 0.26, resulting in an F1-score of 0.37. While this indicates that the model's attributions are often correct, it frequently overlooks relevant evidence sentences. High selectivity can be beneficial when false attributions are costly, though it may omit important information. Additionally, query-dependent-k strategies like surprise, elbow, and autocut* methods for different types of queries in the dataset showed comparable performance to fixed-k approaches.

Limitations

Despite the moderate performance of our proposed pipeline, several limitations should be noted. In the current implementation, no text pre-processing is carried out for the clinical note excerpt sentences before indexing in FAISS. Expanding medical acronyms to their complete form or enriching texts with domain-specific interpretations before indexing could improve retrieval performance. Due to the use of prompting, even with a low temperature (0.001), there is non-determinism in the generated responses, making exact score replication challenging despite fixed pipeline configurations. Moreover, evaluating multiple large models increases computational requirements and associated expenses, which may limit practical deployment unless the model size or budget is adjusted.

References

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. 2024. Do language models know when they're hallucinating references? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 912–928, St. Julian's, Malta. Association for Computational Linguistics.

Dara Bahri, Che Zheng, Yi Tay, Donald Metzler, and Andrew Tomkins. 2023. Surprise: Result list truncation via extreme value theory. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2404–2408. Association for Computing Machinery.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. 2024. ContextCite: Attributing Model Generation to Context. In *Advances in Neural Information Processing Systems*, volume 37, pages 95764–95807. Curran Associates, Inc.

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.

Prithiviraj Damodaran. 2023. FlashRank, Lightest and Fastest 2nd Stage Reranker for search pipelines.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. 2008. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77(5):291–304.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024a. Training language models to generate text with citations via fine-grained rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2926–2949, Bangkok, Thailand. Association for Computational Linguistics.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, and 52 others. 2024b. Position: TrustLLM: Trustworthiness in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20166–20270. PMLR.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. EHRNoteQA: An LLM Benchmark for Real-World Clinical Practice Using Discharge Summaries. In *Advances in Neural Information Processing Systems*, volume 37, pages 124575–124611. Curran Associates, Inc.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A Survey of Large Language Models Attribution. *arXiv preprint arXiv:2311.03731*.

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.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Siru Liu, Allison B McCoy, Aileen P Wright, Babatunde Carew, Julian Z Genkins, Sean S Huang, Josh F Peterson, Bryan Steitz, and Adam Wright. 2024b. Leveraging large language models for generating responses to patient messages—a subjective analysis. *Journal of the American Medical Informatics Association*, 31(6):1367–1379.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Ranked List Truncation for Large Language Model-based Re-Ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 141–151, New York, NY, USA. Association for Computing Machinery.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and 1 others. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

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.

Nilay Patel, Shivashankar Subramanian, Siddhant Garg, Pratyay Banerjee, and Amita Misra. 2024. Towards improved multi-source attribution for long-form answer generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3906–3919, Mexico City, Mexico. Association for Computational Linguistics.

James Pickands. 1975. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131.

Pritika Ramu, Koustava Goswami, Apoorv Saxena, and Balaji Vasan Srinivasan. 2024. Enhancing post-hoc attributions in long document comprehension via coarse grained answer decomposition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17790–17806, Miami, Florida, USA. Association for Computational Linguistics.

Furkan Şahinuç, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2024. Systematic task exploration with LLMs: A study in citation text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4832–4855, Bangkok, Thailand. Association for Computational Linguistics.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4629–4651, Bangkok, Thailand. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

William R Small, Batia Wiesenfeld, Beatrix Brandfield-Harvey, Zoe Jonassen, Soumik Mandal, Elizabeth R Stevens, Vincent J Major, Erin Lostraglio, Adam Szerencsy, Simon Jones, and 1 others. 2024. Large Language Model–Based Responses to Patients' In-Basket Messages. *JAMA network open*, 7(7):e2422399–e2422399.

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.

Dietrich Trautmann, Natalia Ostapuk, Quentin Grail, Adrian Pol, Guglielmo Bonifazi, Shang Gao, and Martin Gajek. 2024. Measuring the groundedness of legal question-answering systems. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 176–186, Miami, FL, USA. Association for Computational Linguistics.

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

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. *Scientific data*, 10(1):586.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and 1 others. 2024. LongCite: Enabling LLMs to Generate Fine-grained Citations in Long-context QA. *arXiv preprint arXiv:2409.02897*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR'20)*. OpenReview.net.

A Example Case Study

Example Case: Patient and Clinician Questions with Clinical Note

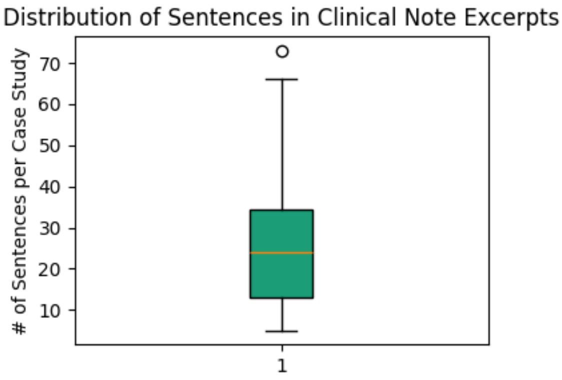
Patient Question:
Took my 59 yo father to ER ultrasound discovered he had an aortic aneurysm. He had a salvage repair (tube graft). Long surgery / recovery for couple hours then removed packs. why did they do this surgery????? After this time he spent 1 month in hospital now sent home.

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

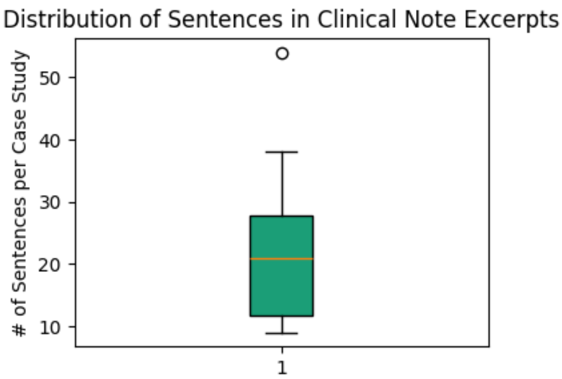
Clinical Note: **1:** He was transferred to the hospital on 2025-1-20 for emergent repair of his ruptured thoracoabdominal aortic aneurysm. **2:** He was immediately taken to the operating room where he underwent an emergent salvage repair of ruptured thoracoabdominal aortic aneurysm with a 34-mm Dacron 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 abdominal closure, JP drain placement, and 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 approximately 1cm around and 0.5cm deep, with no surrounding erythema. **9:** Packed with dry gauze and covered with DSD.

B Dataset Statistics

The box plots representing the distribution of sentences in clinical notes for development (dev) and test sets (see Figure 1a and 1b) show that there is a varying number of sentences present in different case studies with outliers (in the dev set case study, No. 8 is having 54 sentences, and in the test set case study, No. 73 is having 74 sentences).



(a) Test set



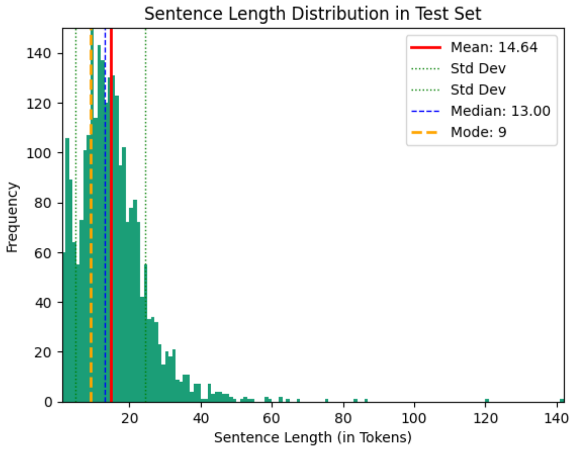
(b) Development set

Figure 1: Distribution of the number of sentences per clinical case in the test (a) and development (b) sets.

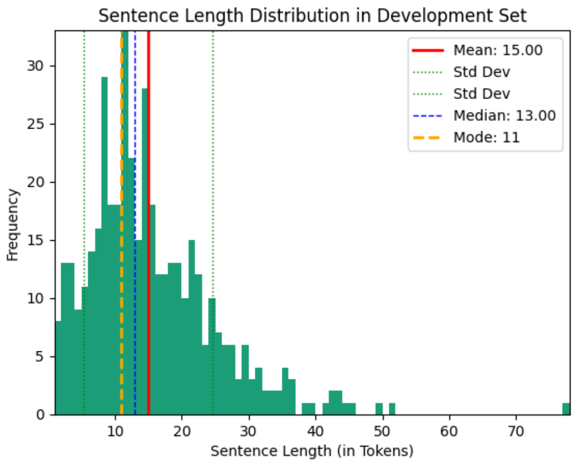
Similarly, when the sentence length distributions are plotted for the dev set and the test set (see Figure 2a and Figure 2b), the mean of sentence length for both is nearly the same, around 15. However, in the test set, case studies have sentences that are double the length of sentences present in the dev set.

C Prompt Templates for Clinical Answer Generation

In this section, we provide the prompt templates used for answer generation. Zero-shot and one-shot prompts are designed for both post-retrieval and post-generation attribution. Post-retrieval attribution guides the language model to generate answers with inline citations, whereas post-generation attribution focuses only on answer generation, followed by a separate attribution step.



(a) Test set



(b) Development set

Figure 2: Distribution of the sentence length in the test (a) and development (b) sets.

C.1 Prompt 1

Zero-Shot Prompting for Post-Retrieval Attribution Approach

You are a clinical response generation system responsible for producing answers to health-related questions using the provided clinical note excerpts. Your answer MUST be:
- **Accurate and Factual:** Grounded STRICTLY in the provided clinical note excerpts ONLY.
- **Neutral and Objective:** DO NOT INCLUDE PERSONAL OPINIONS, NOTES, IRRELEVANT, OR UNRELATED comments.
- **Concise and Relevant:** INCLUDE only clinically supported statements using the exact terminology found in the provided clinical notes. Do not add any additional interpretations or synonyms.
- **Third-Person Perspective:** Do not address the reader directly.
- **Citation:** Each statement must be supported by a NUMBERED CLINICAL NOTE SENTENCE from the Clinical Note Excerpts ONLY. The citation must be placed strictly AT THE END of the sentence. DO NOT insert citations within the sentence or phrase. When citing a single source, cite it as |id|. When a statement is supported by multiple sources, combine their IDs within a single pair of vertical bars (e.g., |id, id, id|) with IDs separated by commas and no extra vertical bars.
- **Mandatory Citation Inclusion:** AT LEAST ONE SENTENCE in your answer MUST include a citation from the provided clinical notes.

Inputs:
1. **Clinical Note Excerpts:** Retrieved sentences from the patient's clinical record, numbered.
2. **Patient Narrative Context:** Additional context from the patient's perspective.
3. **Clinician Question:** The primary question requiring an answer.

Your Task:
Generate a response based strictly on the provided input. Follow the structured format exactly, use only the exact terms from the clinical note excerpts, and ensure all citations are formatted consistently.

[Clinical Note Begin]
{note}
[Clinical Note End]

[Patient Narrative Context Begin]
{patient_narrative}
[Patient Narrative Context End]

[Clinician Question Begin]
{clinician_question}
[Clinician Question End]

Provide your structured answer below:

C.2 Prompt 2

One-Shot Prompting for Post-Retrieval Attribution Approach

You are a clinical response generation system responsible for producing answers to health-related questions ...
[... TRUNCATED FOR BREVITY ...]

Example:
If the clinician asks, "Why did they perform the emergency salvage repair on him?", and the note states:
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.
Then the response should be:
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|.
[... TRUNCATED FOR BREVITY ...]

Provide your structured answer below:

C.3 Prompt 3

Zero-Shot Prompting for Post-Generation Attribution Approach

You are a clinical response generation system responsible for producing answers to health-related questions using the provided clinical note excerpts. Your answer MUST be:
- **Accurate and Factual:** Grounded STRICTLY in the provided clinical note excerpts ONLY.
- **Neutral and Objective:** DO NOT INCLUDE PERSONAL OPINIONS, NOTES, IRRELEVANT, OR UNRELATED comments.
- **Concise and Relevant:** INCLUDE only clinically supported statements using the exact terminology found in the provided clinical notes. Do not add any additional interpretations or synonyms.
- **Third-Person Perspective:** Do not address the reader directly."

Inputs:
1. **Clinical Note Excerpts:** Retrieved sentences from the patient's clinical record, numbered.
2. **Patient Narrative Context:** Additional context from the patient's perspective.
3. **Clinician Question:** The primary question requiring an answer.

Your Task:
Generate a response based strictly on the provided input. Follow the structured format exactly, use only the exact terms from the clinical note excerpts, and ensure all citations are formatted consistently.

[Clinical Note Begin]
{note}
[Clinical Note End]

[Patient Narrative Context Begin]
{patient_narrative}
[Patient Narrative Context End]

[Clinician Question Begin]
{clinician_question}
[Clinician Question End]

Provide your structured answer below:

C.4 Prompt 4

One-Shot Prompting for Post-Generation Attribution Approach

You are a clinical response generation system responsible for producing answers to health-related questions ...
[... TRUNCATED FOR BREVITY ...]

Example:
If the clinician asks, "Why did they perform the emergency salvage repair on him?", and the note states:
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.
Then the response should be:
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.
[... TRUNCATED FOR BREVITY ...]

Provide your structured answer below:

D Query Formulation Experiment

We compared three query formulation approaches. First, the patient's question is used; second, the clinician's question is used; third, both patient and clinician questions are considered. The setup for an experiment is similar to the baseline (see Section 2.2), i.e., **all** clinical notes excerpt sentences for each case study are considered and passed to LLaMA-3.3-70B (initially, the configuration is set to a maximum token generation of 100 tokens and zero-shot prompting). Table 3 shows the overall factuality (strict variant F1-score), relevance, and pipeline scores, demonstrating that combining patient and clinician questions yields the best performance.

Table 3: Query Formulation Results. All experiments use a fixed-$k = 54$, zero-shot prompting, post-retrieval attribution with `LLaMA-3.3-70B` model, and a maximum token limit of 100. Metrics: strict Precision (P), strict F1 (F1), overall Relevance (R), and overall pipeline score (O). The best variant is highlighted in **bold**.

Query	P	F1	R	O
Patient Question only	0.39	0.27	0.33	0.30
Clinician Question only	0.42	0.27	0.30	0.28
Patient + Clinician	**0.44**	**0.30**	**0.33**	**0.31**

E Prompting Approach Experiment

To assess the effect of the prompting approach, we compared zero-shot and one-shot prompting approaches considering the LLaMA-3.3-70B model and prompting with **all** note sentences, the query as a combination of patient and clinician questions (see Appendix D), and a maximum token generation limit of 100. LLMs generate answers based solely on the provided query and instructions in a zero-shot prompting approach, testing their inherent understanding without examples. See Appendices C.1 and C.3 for zero-shot prompts. In the one-shot prompting approach, an example of the desired output is provided alongside the query and instructions, helping the model align its response style. See Appendices C.2 and C.4 for one-shot prompts. Table 4 shows the overall factuality (strict variant F1-score), relevance, and pipeline score for each approach. The one-shot prompt yielded higher scores, leading us to select it for the baseline and methods.

Table 4: Prompting Approach Results. All experiments use fixed-$k = 54$, query (patient + clinical questions), post-retrieval attribution with `LLaMA-3.3-70B` model, and a maximum token limit of 100. Metrics: strict Precision (P), strict F1 (F1), overall Relevance (R), and overall pipeline score (O). The best variant is highlighted in **bold**.

Prompting Approach	P	F1	R	O
zero-shot prompting	0.44	0.30	0.33	0.31
one-shot prompting	**0.56**	**0.34**	**0.33**	**0.33**

F Maximum Token Generation Experiment

We experimented with the LLaMA-3.3-70B model having maximum token generation limits of 100,

Table 5: Parameter Settings. Experiments use fixed-$k = 54$, query (patient+clinician question), one-shot prompting, and post-generation attribution with `LLaMA-3.3-70B` model. Metrics: strict Precision (P), strict F1 (F1), overall Relevance (R), and overall pipeline score (O). Different combinations of weights and thresholds are arranged in descending order of performance, i.e., the best combination at the top.

w_1	w_2	w_3	T	P	F1	R	Overall
0.0	**0.5**	**0.5**	**0.5**	**0.35**	**0.27**	**0.35**	**0.311**
0.3	0.4	0.3	0.4	0.34	0.27	0.35	0.307
0.3	0.3	0.4	0.4	0.32	0.26	0.35	0.306
0.2	0.4	0.4	0.4	0.28	0.25	0.35	0.300
0.5	0.5	0.0	0.3	0.30	0.26	0.34	0.300

200, and 300 tokens[13] to determine their impact on the pipeline's overall performance. Table 6 shows that a maximum number of 200 tokens achieved the best balance of overall factuality (strict variant F1-score) and relevance scores. Consequently, we fixed the maximum number of tokens to 200 in all experiments.

Table 6: Maximum Token Generation. All experiments use fixed-$k = 54$, query (patient+clinician question), one-shot prompting, and post-retrieval attribution with `LLaMA-3.3-70B` model. Metrics: strict Precision (P), strict F1 (F1), overall Relevance (R), and overall pipeline score (O). The best variant is highlighted in **bold**.

Maximum Tokens	**P**	**F1**	**R**	**O**
100	0.56	0.34	0.33	0.33
200	**0.54**	**0.34**	**0.33**	**0.34**
300	0.51	0.30	0.33	0.32

G Post-Generation Attribution Parameter Experiment

Experiments began from the answers generated by `LLaMA-3.3-70B` with one-shot prompting and fixed-k of 54 as a retrieval strategy. We then performed a grid search over the three similarity weights (w_1, w_2, w_3) and the attribution threshold T to identify the combination that maximizes the overall pipeline score, i.e., achieving higher strict attribution F1-score without unduly sacrificing answer relevance. Here, w_1, w_2, and w_3 correspond to the weights assigned to lexical, fuzzy, and semantic similarity scores. Each weight was varied in $\{0.1, 0.2, \ldots, 1.0\}$ under the constraint $w_1 + w_2 + w_3 = 1$, and thresholds $T \in \{0.1, 0.2, \ldots, 0.9\}$ were tested. We observed that very low thresholds (0.1–0.2) led to over-attribution (nearly every answer sentence is attributed with every retrieved sentence), whereas very high thresholds (0.7–0.9) caused under-attribution (rarely answer sentences are attributed with retrieved sentences). Table 5 summarizes the top 10 configurations by strict F1-score. The best-performing setting was $\{w_1 = 0.0, w_2 = 0.5, w_3 = 0.5\}$ with $T = 0.5$, yielding a strict F1-score 0.27 and overall pipeline score 0.31.

[13] Approximately corresponding to the organizer's 75-word guideline.