Loyola at ArchEHR-QA 2025: Unsupervised Methods for Text Attribution with Attention and Clustering

Rohan Sethi
Loyola University Chicago
rsethi1@luc.edu

Timothy A. Miller
Boston Children's Hospital, Harvard Medical School
timothy.miller@childrens.harvard.edu

Majid Afshar
University of Wisconsin-Madison
mafshar@medicine.wisc.edu

Dmitriy Dligach
Loyola University Chicago
dd@cs.luc.edu

Abstract

The increasing volume of patient messages via electronic health record (EHR) portals has contributed significantly to clinician workload. Automating responses to these messages can help alleviate this burden, but it is essential to ensure that the generated responses are grounded in accurate clinical evidence. As part of the ArchEHR-QA 2025 BioNLP ACL shared task, we explore unsupervised methods for generating patient question responses that are both contextually accurate and evidence-backed. We investigate three novel approaches: zero-shot prompting, clustering-based evidence selection, and attention-based evidence attribution, along with a hybrid model that combines clustering and attention. Our methods do not require model fine-tuning and leverage the inherent structure of the input data to identify the most relevant supporting evidence from clinical notes. Our best-performing approach, which integrates clustering and attention, demonstrates a substantial improvement in factuality over baseline zero-shot methods, highlighting the potential of unsupervised strategies for enhancing the clinical utility of large language models in EHR contexts.

1 Introduction

Electronic health record (EHR) systems have improved physicians' ability to document and track patient care over time. They also facilitate digital communication, allowing patients to engage with their health goals beyond in-person visits. However, the rise in patient messaging has unintentionally added to clinician workload (National Academies of Sciences et al., 2019).

Large language models (LLMs) have been proposed as tools to automatically answer patient questions. However, mere generation is not sufficient; responses must be grounded in clinical evidence from patients' medical records to ensure accuracy and reliability (Lin et al., 2003). The ArchEHR-QA 2025 BioNLP ACL shared task (Soni and Demner-Fushman, 2025b) aims to develop systems that can generate such grounded answers using information extracted from EHRs. Thus, the task is to generate an answer to a patient's question and include the sentences (or sentence identifiers) from the source note as supporting evidence for the answer.

The problem of evidence attribution has received much attention recently, and can be categorized as follows: direct LLM attribution, post-retrieval generation, and post-generation attribution. Some approaches prompt the LLM to directly generate attribution within its responses. However, (Zuccon et al., 2023) investigates ChatGPT's ability to attribute directly using prompting strategies and found that the attributions was partially correct around 50% of the time and only present 14% of the time demonstrating its unreliability. Other approaches attempt to retrieve relevant external information and prompt an LLM to incorporate said information during generation. However, citations for these approaches were present only 50% of the time (Gao et al., 2023). Finally, (Liu et al., 2023) investigates the quality of citations generated by mainstream generative search engines that incorporate citations post-generation. It was found that only 51.5% of generated sentences were fully supported and that 74% of the citations supported their associated sentences (Liu et al., 2023). Clearly, current methods to attribute are lacking in consistency and relevance of LLM text attribution.

As with many clinical machine learning tasks, this challenge provides limited data - only 20 training and development questions with corresponding medical records. To address the data scarcity, we propose two novel unsupervised methods that do not require fine-tuning or alignment of LLMs. This paper examines two approaches individually and in combination. The first uses clustering to identify the most relevant clinical evidence for a given question, narrowing the context for LLM input. The

BioNLP 2025 Shared Tasks, pages 22–26

second employs an attention-averaging augmented generation method, where the LLM generates a response freely, and attention weights are used post hoc to attribute evidence sources. We also evaluate a combined approach and compare all methods against a baseline that prompts the LLM without any augmentation.

2 Methods

2.1 Dataset

The dataset (Soni and Demner-Fushman, 2025a) is adapted from the Medical Information Mart for Intensive Care (MIMIC) corpus (Johnson et al., 2016) by the organizers of the ArchEHR-QA 2025 BioNLP ACL shared task (Soni and Demner-Fushman, 2025b). It consists of patient-inspired questions paired with relevant clinical note excerpts from MIMIC, forming "cases." Each excerpt is pre-annotated, with sentences labeled as "essential," "supplementary," or "not relevant" for answering the question. Sentences are numbered to serve as citations in generated responses. A physician-paraphrased version of the patient's question is also provided. The development set includes 20 cases, while the test set contains 100.

2.2 Evaluation

Evaluation is based on two metrics - Factuality and Relevance - and their average. Factuality is measured using precision, recall, and F1 score between the system-selected citations and the gold-standard "essential" evidence. Relevance compares the generated response to a paragraph combining the question and essential evidence text, using BLEU, ROUGE, SARI, BERTScore, AlignScore, and MEDCON. Evaluation scripts are provided by the challenge organizers.

2.3 Method details

We propose three methods for answer generation and evidence attribution, and compare them to a baseline provided by the shared task organizers, which prompts an LLM to answer questions. The first method uses clustering to identify relevant citations based on sentence groupings. The second leverages transformer attention to attribute evidence to each generated sentence. The third combines both approaches, using clustering to guide attention-based attribution. Details of each method are provided below. All experiments were run on a NVIDIA RTX A6000 GPU. We release the code that is necessary to reproduce our experiments[1].

Zero and few-shot baselines: The baseline was performed by the organizers of the ArchEHR competition that involves prompting LLaMa 3.3 70B (AI@Meta, 2024) on the test set in a zero-shot fashion. The model is prompted to generate answers that included citations. If a response was invalid (e.g., exceeding the word limit or lacking valid citations), the prompt was retried up to five times to obtain a valid output. Additionally, we explored including multiple examples in the prompt (few-shot) but this lead to significant performance degradation on the development set.

Clustering-based method: First, every clinical note sentence and the concatenated patient and physician versions of the question are converted into embeddings. Embeddings are obtained using an encoder LLM via HuggingFace's transformers feature extraction API (Wolf et al., 2020). Embeddings are provided per token, so the token embeddings for each sentence are averaged to get an overall embedding for the given sentence. These embeddings are then clustered into two clusters using the agglomerative clustering algorithm from sci-kit learn (Pedregosa et al., 2011). The clusters are then parsed to identify the cluster containing the question embedding vector. The clinical sentences that are a part of this cluster are assumed to contain the most relevant evidence to answer the patient's question. These clinical sentences are used as input to the LLM prompt, which is prompted to answer the patient question given the selected context. Post-generation, the clinical sentences utilized are cited at the end of the LLM response without precisely attributing each output sentence to a clinical sentence. This is unlike the attention-based and hybrid approaches which precisely cite clinical note segments to each output sentence. An example is included in the prompt to demonstrate to the model how detailed its response should be without restrictions on formatting responses.

Attention-based model: This method leverages transformer attention scores to attribute generated text to specific sentences in the source clinical note. We hypothesize that the average attention between generated output and input sentences can serve as a signal for source attribution. All questions and evidence entries are input to an LLM. An example is provided in the prompt to demonstrate to the

[1]https://github.com/rsethi21/loyola_archehr_2025.git

model how detailed its response should be. This example, however, does not give restrictions on how responses to be formatted. Post-generation, attention outputs are analyzed. For each output sentence, an average attention score is computed with regards to each evidence entry; i.e. if there were n evidence entries, there will be n computed average attention scores for each output. To obtain attention scores for averaging, we parse the attention matrix from the LLM after determining the token indices of output sentences and each evidence entry. Details on how the indexed attention matrix is utilized to compute average scores can be found in the source code.

All computed evidence entry attention scores for each output are converted to z-scores, and entry scores exceeding a threshold are considered supporting evidence, which is then appended to the corresponding output sentence. The z-score selection enables selection of only the most significantly attended evidence entries or alternatively no citations if all z-scores are below a threshold, which makes this attribution factually robust. This process is repeated for all output sentences in the LLM response. Key hyperparameters include the LLM model, prompt format, z-score threshold, and chosen attention layers.

Hybrid model: The final method combines clustering and attention-based approaches. The LLM is first prompted with evidence selected via clustering, and its output is then processed using the attention-based attribution workflow. This hybrid method tests whether clustering can guide the LLM's attention toward the most relevant evidence, potentially improving the identification of essential information compared to using either method alone.

Methods not included in the final submission: Other methods were experimented with early in the competition, including retrieval augmented generation (RAG), few-shot prompting, encoder-based evidence selection, supervised-fine tuning, selection of evidence based on similarity to output post-generation using BERTScore, and others. However, the best performing methods were finalized for submission using the development set and described above.

2.4 Experiments

All models are implemented using HuggingFace (Wolf et al., 2020) and PyTorch (Paszke et al., 2019). We use Llama-3.1-8B-Instruct (AI@Meta, 2024) and we conduct all experiments on 8 NVIDIA RTX A6000 GPUs. All hyperparameters are tuned on the development set.

For the clustering approach, the dmis/biobert v1.1 (Lee et al., 2020) was selected against other embedding models. Agglomerative clustering with number of clusters of 2 was selected from SkLearn's clustering module (Pedregosa et al., 2011). The options for clustering algorithms included KMeans, Agglomerative, and DBSCAN. Number of clusters was varied between 2 and 3 representing the categories that the evidence entries were labeled as (essential vs not relevant or essential vs supplementary vs not relevant).

For the attention approach, selection of attention layers was treated as a hyperparameter. The attention layers kept varied by case, where attention layer outputs were compared sequentially using cosine similarity and only the attention output layers that differed the most from the previous attention layer was kept for averaging. The hyperparameter was whether to perform this selection or not. For the development set, dropping attention layers was selected for and applied to the test cases. The selected for z-score threshold was 0 from the following options 1.64 (average attention score significantly greater than 95% of other attention scores), 1 (significantly greater than 85%), and 0 (significantly greater than 50%).

All methods are evaluated on the development set using the scoring script provided by the organizers of the competition.

3 Results and Discussion

Experiment	Factuality	Relevance	Average
Zero-Shot	43.10	28.70	35.90
Clustering	50.56	32.38	41.47
Attention	54.11	31.81	42.96
Clustering + Attention	58.64	33.37	46.00

Table 1: Development set overall factuality, overall relevance, and overall scores for all methods. Zero-shot is the baseline approach attempted by the organizers of the competition.

The results of our performance evaluation on the development set are presented in Table 1. The best-performing method is the hybrid approach, Clustering combined with Attention, which improves the factuality score from a baseline of 43.10 to 58.64. The attention-based method alone achieves a score of 54.11, while the clustering-only method yields 50.56. These results suggest that the model's

Experiment	Factuality	Relevance	Average
Zero-Shot	33.60	27.80	30.70
Clustering + Attention	57.35	30.36	43.85

Table 2: Test set overall factuality, overall relevance, and overall scores for best method and zero-shot. Zero-shot is the baseline approach implemented by the organizers of the competition.

attention matrix can effectively highlight the information the LLM prioritizes when generating each sentence, contributing to a nearly 10-point increase in factuality. By leveraging attention, LLMs can generate more accurate outputs without relying on complex formatting or explicit instructions, while also enabling real-time evidence integration during generation.

Additionally, combining clustering with attention further improved the factuality score by 4 points over using attention alone. This indicates that selecting relevant evidence through clustering before passing it to the LLM helps the model focus more effectively on the most pertinent information when answering patient questions, leading to higher factuality.

In terms of relevance, the greatest improvement over the zero-shot baseline came from combining clustering and attention, resulting in a nearly 5-point gain. This likely stems from more accurate evidence selection.

The best-performing approach, clustering combined with attention, was evaluated on the test set and compared to the organizer's zero-shot baseline. It maintained similar average scores for both factuality and relevance, showing no significant performance drop and achieving comparable gains over the baseline as seen on the development set. Notably, this unsupervised method using an 8B LLM outperformed a 70B LLM, offering substantial savings in computational cost, time, and training resources. Curating clinically oriented training datasets is both time-consuming and resource-intensive, making them difficult to obtain. Our results demonstrate that unsupervised methods can effectively enhance the factuality of LLM-generated responses in clinical settings.

4 Conclusion

Automating responses to patient questions using EHR data holds significant potential for reducing clinician workload and improving patient care. In this work, we demonstrated that integrating unsupervised approaches like clustering and attention-based evidence attribution with large language models (LLMs) can significantly enhance the factuality and relevance of generated responses without requiring extensive model fine-tuning or alignment. Our hybrid method, combining clustering and attention, outperformed traditional zero-shot baselines, highlighting the value of leveraging context structuring and attention analysis for more accurate clinical responses. Importantly, our findings show that relatively small LLMs (8B parameters) can outperform much larger models (70B parameters) when appropriately guided, offering substantial cost and efficiency advantages in real-world clinical applications. Future work could further refine these methods by incorporating more sophisticated context selection strategies, leveraging multimodal data, and exploring more interpretable attention mechanisms to ensure even higher levels of clinical trustworthiness and reliability.

5 Limitations

Most experiments were performed utilizing LLMs with 8B parameters or less due to memory constraints. Furthermore, only 20 development / training examples were provided for experimentation. Although these examples labeled the evidence entries that were essential to incorporate in the answer the experimented approaches generate, there were no associated example answer outputs.

Acknowledgments

Research reported in this publication was supported by National Institutes of Health under Awards R01LM012973 and 1R01DA051464. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

AI@Meta. 2024. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md. Accessed: 2025-05-06.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. 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.

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. *Scientific Data*, 3(1):160035.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jimmy Lin, Dennis Quan, Vineet Sinha, Krzysztof Bakshi, David Huynh, Boris Katz, and David R. Karger. 2003. What makes a good answer? the role of context in question answering. In *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2003)*. Springer.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *ArXiv*, abs/2304.09848.

Engineering National Academies of Sciences, Medicine, National Academy of Medicine, and Committee on Systems Approaches to Improve Patient Care by Supporting Clinician Well-Being. 2019. *Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being*. National Academies Press. Accessed: 2025-05-06.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

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 24rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Guido Zuccon, Bevan Koopman, and Razia Shaik. 2023. Chatgpt hallucinates when attributing answers. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '23, page 46–51, New York, NY, USA. Association for Computing Machinery.