Evaluating the Synergistic Impact of Fine-Tuning and Retrieval-Augmented Encoding on Enhancing Appendicitis Diagnosis from Limited HPI Notes

Lubna Mahmoud Abu Zohair[1*], **Hind Zantout**[2]

[1]Independent Researcher

[2]School of Mathematical and Computer Science, Heriot-Watt University

Dubai, United Arab Emirates

[*]lubna.abuzohair@gmail.com

Abstract

This study explores the impact of fine-tuning combined with a retrieval-augmented encoding approach on encoder language model-generated embeddings for appendicitis diagnosis tasks using patients' History of Present Illness notes, leading to significantly enhanced diagnostic performance.

Background, Problem, and Objective

In clinical settings, particularly when dealing with limited data, enhancing language model inference for diagnosing rare or less common conditions in resource and computational constrained environments presents a critical challenge. Published clinical data is often scarce, especially for rare diseases or conditions, and is rarely encountered by existing general or medical-specific pre-trained language models due to privacy constraints (Pieper et al. 2024; Safonova et al. 2023; Schick and Schütze 2021; YU et al. 2023). If there are no privacy concerns, zero-shot current state-of-the-art large language models (LLMs) are considered the ultimate solution for common knowledge domain-related problems. Retrieval-Augmented Generation (RAG) can complement LLM limitations in less common topics without the need for fine-tuning and help in mitigating hallucinations (Soudani et al. 2024). However, for the current research about clinical scenarios that this research is targeting, the use of relevant, affordable (with less than billions number of parameters) pre-trained language models, particularly locally downloadable ones, is often viewed as the ultimate solution. Also, since decoder-based models are prone to hallucinations, encoder-only models will be the right candidate for model inference tasks, like medical diagnosis. Nevertheless, these models, on their own, remain insufficient when addressing new medical conditions with insufficient data. A nearly relevant example of this challenge arises when trying to diagnose appendicitis from other abdominal disease (with differential diagnosis) based solely on a limited set of patient's history of present illness (HPI) notes. While appendicitis is not a rare disease, the clinical notes of a patient are often private and inaccessible publicly to current state of the art LLMs. Besides, clinical notes for rare diseases cannot be obtained easily attainable from clinical providers, making Appendicitis HPI notes a compelling case for assessing the proposed approach for maximizing the inference of affordable encoder language models in limited dataset condition. Fine-tuning was proven to be applicable for improving the relevance and accuracy of model inferences. However, it requires adequate and representative datasets; otherwise, it may struggle to converge effectively in domains with limited data, as insufficient or biased learned weights can result from the scarcity of examples. Researchers found RAG adoption as a regularization technique can improve the decoder only model generation, especially when combined with model fine-tuning, enhancing the quality of language model generation for small datasets (Mallen et al. 2023; Soudani et al. 2024). However, these studies explored the utility of combining both approaches for improving text generation tasks, not for augmenting the quality of the embeddings generated by encoder-only models which are characterized by their enhanced inference capabilities. But in the case of this research the generation part of RAG is replaced by encoding (RAE). The knowledge resource in RAE will be appendicitis notes of previously diagnosed patients. The rationale for this is the fact that in rare conditions medical expertise is not solely derived from understanding pathophysiological mechanisms but also from developing illness scripts, memories, or previous expertise through exposure to real and similar patient cases (Brooks et al. 1991; Schmidt et al. 1992). Given that context, will fine-tuning medical specific pre-trained model, employing RAE, or combining both fine-tuning and RAE yield the best results in improving the quality of language model inferences? The combination of fine-tuning and RAG holds the potential to leverage the strengths of both approaches. Fine-tuning can refine the model's understanding of appendicitis in the context of limited HPI notes, while RAG can augment this understanding by providing additional context from external sources of clinical knowledge. This research evaluates these

approaches by enhancing the diagnostic inference of appendicitis using HPI notes. Focusing on this specific medical condition, the study aims to determine which method, or combination of methods, provides the most accurate and reliable model outputs when working with small clinical datasets. Through this evaluation, we aim to provide insights into the optimal strategy for improving model inference accuracy in clinical scenarios where data and resources are limited, and diagnostic precision is critical.

Data and Methods

Appendicitis HPI notes and diagnoses were sourced from the MIMIC-IV-EXT database (Hager et al. 2024), Under a Data License that prohibits processing data using external API-based language models (such as OpenAI, Google Gemini, Meta, etc.), only affordable, open-access encoder models available for direct download were used. The key dataset preprocessing phases include the exclusion of notes with multiple abdominal diagnoses (differential diagnoses). Furthermore, Appendicitis was treated as one class, with other abdominal diseases as another, to mitigate classes imbalances, resulting in 2331 different patient notes. Further filtering removed notes exceeding 256 tokens, leaving 2203 notes for the subsequent analysis. Then, the dataset was split into two-thirds for training and one-third split evenly for testing and a RAG set. Three-fold cross-validation was used to ensure robust model evaluation. Initial analysis was run for multiple encoder models with classification heads for the diagnosis of appendicitis from HPI notes, and the top performing models were bio-formers/bioformer-8L' and 'bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12' (Fang et al. 2023; Peng et al. 2019). Naïve RAG structure was used but replacing the generator model part with the selected encoder models, hence the Encoding part of the name RAE. For diagnosis, RAE retrieved similar notes from the RAG set, which were then concatenated with tested notes. The FAISS index and the 'BAAI/bge-small-en-v1.5' embedding model were widely employed and used in this research to index, store, and retrieve relevant notes within the RAE structure (Arslan et al. 2024; Fan et al., 2024; Langchain 2024). Model performance was compared with and without RAE queries, before and after model fine-tuning on a binary classification problem to identify appendicitis (encoded by 1) other abdominal diseases (encoded by 0). The classification accuracy, precision, F1-score, and recall were employed as performance measures. Last, statistical paired t-test was employed to determine whether there is a statistically significant difference (and increase) in the mean scores after fine tuning with RAE, with p-value threshold set to 0.05. The tested null hypothesis is the synergy of fine-tuning, and the RAE approach does not lead to a notable improvement in appendicitis diagnostic performance.

Results

BlueBERT and Bioformer-8L models diagnosing performance were evaluated across various test conditions, with and without models fine-tuning, with and without the application of RAE, as presented in Table 1. In the fine-tuned scenario, BlueBERT showed significant improvements, with accuracy reaching 0.8883, precision 0.8921, recall 0.8883, and F1 score 0.8886. However, when combined with the RAE application, the fine-tuned BlueBERT model demonstrated further enhancement, achieving an accuracy of 0.8919, precision 0.8922, recall 0.8919, and F1 score of 0.892. Similarly, Bioformer-8L reached an accuracy of 0.8901, precision 0.8911, recall 0.8901, and F1 score 0.8902, after being fine-tuned, with additional 1% increase in the model diagnosing performance, yielding the highest accuracy and precision readings as 90.37% and 90.39%, respectively. Lastly, a paired t-test indicated a significant improvement in the values of performance metrics for the HPI model after applying fine-tuning and RAE, with a two-tailed p-value of 0.00597, rejecting the null hypothesis by confirming that this combined approach resulted in a statistically significant positive enhancement.

Conclusion

Fine-tuning, especially when combined with the RAE approach, can significantly maximize the encoder language model embedding quality and quantity when dealing with limited clinical notes, making it sufficient for optimizing the model's classification task, when such models are combined with a classification head. This reaffirms the findings of Mallen et al. (2023) and Soudani et al. (2024) that such an approach is also necessary for enhancing encoder models, especially when dealing with small datasets. Notably, applying RAE improved performance consistency in the low-parameter Bioformer-8L model, especially in zero-shot tests, enhancing true positive detection and reducing false results despite class imbalance and limited data (Batista et al. 2004).

Notes and Test conditions	Models Performance							
	BlueBERT				Bioformer-8L			
	Average Accuracy	Average Precision	Average Recall	Average F1 Score	Average Accuracy	Average Precision	Average Recall	Average F1 Score
HPI (Zero Shot)	0.4687	0.4706	0.4687	0.4692	0.5441	0.3491	0.5441	0.4015
HPI (Zero Shot with RAE)	0.4932	0.5062	0.4932	0.4947	0.5459	0.5183	0.5459	0.4952
HPI (fine-tuned)	0.8883	0.8921	0.8883	0.8886	0.8901	0.8911	0.8901	0.8902
HPI (fine-tuned with RAE)	0.8919	0.8922	0.8919	0.892	0.9037	0.9039	0.9037	0.9037

Table 1: Model Performance Across Multiple HPI-Tested Conditions.

References

Arslan, M.; Ghanem, H.; Munawar, S.; and Cruz, C. 2024. A Survey on RAG with LLMs. *Procedia Computer Science* 246: 3781–3790. https://doi.org/10.1016/j.procs.2024.09.178.

Batista, G. E. A. P. A.; Prati, R. C.; and Monard, M. C. 2004. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explorations Newsletter* 6(1): 20–29. https://doi.org/10.1145/1007730.1007735.

Brooks, L. R.; Norman, G. R.; and Allen, S. W. 1991. Role of Specific Similarity in a Medical Diagnostic Task. *Journal of Experimental Psychology: General* 120(3): 278–287. https://doi.org/10.1037/0096-3445.120.3.278.

Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; and Li, Q. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*: 6491–6501. https://doi.org/10.1145/3637528.3671470.

Fang, L.; Chen, Q.; Wei, C.-H.; Lu, Z.; and Wang, K. 2023. Bioformer: An Efficient Transformer Language Model for Biomedical Text Mining.

Hager, P.; Jungmann, F.; and Rueckert, D. 2024. MIMIC-IV-Ext Clinical Decision Making: A MIM-IC-IV Derived Dataset for Evaluation of Large Language Models on the Task of Clinical Decision Making for Abdominal Pathologies (version 1.1). *PhysioNet*, July 8.

Langchain. 2024. Faiss. https://python.langchain.com/docs/integrations/vectorstores/faiss/.

Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; and Hajishirzi, H. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*: 9802–9822. https://doi.org/10.18653/v1/2023.acl-long.546.

Peng, Y.; Yan, S.; and Lu, Z. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *Proceedings of the 18th BioNLP Workshop and Shared Task*: 58–65. https://doi.org/10.18653/v1/W19-5006.

Pieper, T.; Ballout, M.; Krumnack, U.; Heidemann, G.; and Kühnberger, K.-U. 2024. Enhancing Small Language Models via ChatGPT and Dataset Augmentation. In *Proceedings of the [Conference Name]*: 269–279. https://doi.org/10.1007/978-3-031-70242-6_26.

Safonova, A.; Ghazaryan, G.; Stiller, S.; Main-Knorn, M.; Nendel, C.; and Ryo, M. 2023. Ten Deep Learning Techniques to Address Small Data Problems with Remote Sensing. *International Journal of Applied Earth Observation and Geoinformation* 125: 103569. https://doi.org/10.1016/j.jag.2023.103569.

Schick, T.; and Schütze, H. 2021. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*: 2339–2352. https://doi.org/10.18653/v1/2021.naacl-main.185.

Schmidt, H. G.; Norman, G. R.; and Boshuizen, H. P. 1992. A Cognitive Perspective on Medical Expertise: Theory and Implication. *Acad Med* 65(10): 611–632.

Soudani, H.; Kanoulas, E.; and Hasibi, F. 2024. Fine-Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge. *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*: 12–22. https://doi.org/10.1145/3673791.3698415.

Yu, H.; Guo, P.; and Sano, A. 2023. Zero-Shot ECG Diagnosis with Large Language Models and Retrieval-Augmented Generation. *Machine Learning for Health (ML4H)*: 650–663.