Grenze International Journal of Engineering and Technology, June Issue

Comparative Evaluation of Embedding Models in Retrieval-Augmented Medical Chatbots

Suman Ghorai[1] and Prashanth K[2]
[1-2]Department of Master of Computer Applications, RV College of Engineering, Bengaluru
Email: sumanghorai.mca24@rvce.edu.in, prashanthk@rvce.edu.in

***Abstract*— This paper aims to conduct a comparative evaluation of MiniLM, BERT, and BioBERT embeddings in the context of Retrieval-Augmented Generation (RAG). The experiments were conducted using PubMedQA and MedMCQA as the benchmark and FAISS as the retrieval system. The experiments showed that BioBERT emerged as a winner in terms of accuracy (0.67) and F1 score (0.576), which is a testament to the effectiveness of domain-specific pretraining in medical domains. Although MiniLM was slightly less accurate, it had significantly lower latency (8.97 ms), which puts it at the forefront as the most efficient model in conversational AI. The retrieval latency was low for all three models, and generation latency was constant at 0.9 s, which confirms the fact that retrieval efficiency is impacted only by the choice of embeddings and not generation synthesis. This paper has shown that there is a trade-off between accuracy and efficiency in medical AI. The experiments are to be extended to multiple vector databases and multi-evidence retrieval to optimize user trust in medical AI.**

***Index Terms*— Biomedical NLP, Retrieval-Augmented Generation, Embedding Models, Medical Chatbots, BioBERT, FAISS.**

I. INTRODUCTION

Artificial Intelligence (AI) and Natural Language Processing (NLP) have revolutionized the development of intelligent question answering systems and chatbots. Previously, the development of intelligent question answering systems and chatbots relied on rule-based approaches and keyword matching, which were useful for information retrieval systems.

However, these approaches were not useful for generating natural responses. The development of deep learning and transformer-based models such as BERT and GPT has revolutionized the development of intelligent question answering systems and chatbots, as these models can generate natural responses and understand the context of the conversation [5], [11].

However, these models are not effective in the medical field, as they cannot interpret medical and scientific terminologies. To overcome these limitations, models such as BioBERT and SciBERT were developed, which are effective in the medical field and can be used for question answering systems and chatbots [2], [3], [6]. Yet, despite all these advances, language models are limited by their training data, which can cause "hallucinations" and outdated information. In healthcare, such inaccuracies can be critical. Retrieval-based models, such as dense passage retrieval, have been introduced to address such shortcomings by basing responses on external knowledge sources [8], [9], [10].

Grenze ID: 01.GIJET.12.2.163_24

II. Related Work

With recent progress in Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies, conversational systems and Medical Question Answering have improved significantly. For example, Transformer-based Large Language Models (LLMs) like BERT [11] and GPT [5] have shown effectiveness in understanding complex language patterns and generating clear responses. However, there is still room to improve AI-based healthcare services. Issues such as incorrect outputs, outdated knowledge, and limited understanding of complex areas remain challenges [17], [18].

To overcome these drawbacks of existing technologies, Retrieval-Augmented Generation (RAG) has proven to be an effective solution. RAG-based systems combine document retrieval with generative models to enable AI systems to respond to queries with knowledge from external sources [1], [7]. Recent advances in biomedical RAG-based systems have shown substantial improvements. For instance, Garg et al. [14] developed a QA system on PubMed using MiniLM Embeddings and FAISS Vector Search. Zhang et al. [13] introduced MedBioRAG, showing substantial improvements over existing systems on PubMedQA and MedQA datasets. Techniques like Dense Passage Retrieval (DPR) [9] and ColBERT [8] improve the efficiency of AI-based systems.

Specific models such as BioBERT [2] and SciBERT [6] have consistently outperformed other general models on various natural language processing tasks such as entity recognition, relation extraction, and even medical question answering [3]. More advanced frameworks such as MEGA-RAG [15] and Self-MedRAG [16] have successfully implemented multi-evidence-based retrieval and self-verification to decrease hallucination occurrences. Furthermore, clinical LLMs have proven to be able to recall medical knowledge, thus aiding in decision-making and healthcare-related information retrieval [17], [18].

However, all existing studies have only implemented a single model or infrastructure. There has been no comparative study on various models or vector databases using a unified RAG framework. Furthermore, all existing studies have only focused on accuracy without considering other essential factors such as latency, scalability, and trust [19], [20]. This study aims to fill this research gap by carrying out a comparative analysis of various models and infrastructures to determine the best configuration to create a reliable medical chatbot system.

III. Research Gap and Problem Statement

Although significant improvements have been made in medical question answering and conversational AI, reliability and accuracy are still major issues to be addressed. For example, LLMs usually provide responses based only on their own training data. [5][11][17] This can lead to incorrect information and outdated content. This issue is especially concerning in the medical field. A Retrieval Augmented Generation (RAG) model has been introduced to help solve this problem by using outside sources of knowledge. [1][7] However, there has been limited research on using multiple embedding models like MiniLM, BERT, and BioBERT within the same RAG model. This approach can significantly affect semantic representation and retrieval accuracy. [2][3][6][13][14] Another area that lacks investigation is retrieval infrastructure. Several systems depend on a single vector database, such as FAISS or Pinecone, without addressing differences in latency, scalability, and efficiency. This is particularly important for biomedical literature, which is vast. Therefore, it is essential to consider how infrastructure impacts the overall responsiveness of the system and the user experience. There is limited emphasis placed on efficiency, scalability, and user trust in terms of accuracy in existing evaluation methods. Assessments of user-centered clarity and reliability of AI-generated medical responses are limited.

In other words, the research gap identified is the absence of a comprehensive comparative evaluation for embedding models and retrieval infrastructures for RAG-based medical chatbots. This research looks at MiniLM, BERT, and BioBERT embedding models. It also reviews FAISS and Pinecone vector databases using PubMedQA [4]. The goal is to improve medical AI systems that are precise, effective, and trustworthy for retrieving healthcare information and supporting decision-making.

IV. Methodology

A. Dataset and Preprocessing

The experiments used the PubMedQA and MedMCQA dataset [4] and the derived corpus, which included text contexts and metadata (processed_chunks.json). Each chunk of text was treated as a knowledge unit, and no further label cleaning as done. A small set of yes/no/maybe questions (N = 5) was used for evaluation.

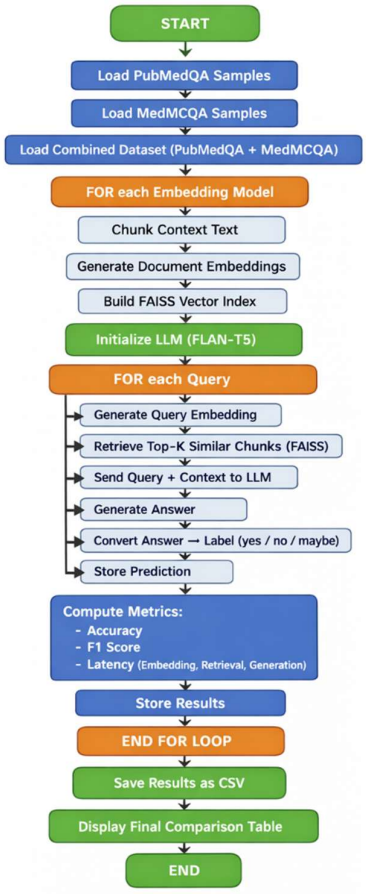

Fig.1: Flowchart of the Experimental Pipeline

B. Embedding Model Comparison

Three embedding models were evaluated:

- sentence-transformers/all-MiniLM-L6-v2
- sentence-transformers/bert-base-nli-mean-tokens
- pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb

Embeddings were computed using SentenceTransformer.encode(...). FAISS indexing was implemented via faiss. IndexFlatL2(dim) for retrieval.

C. Retrieval Pipeline

For each input query, the pipeline encoded the query, conducted top k retrieval (k=5), and aggregated the retrieved contexts into prompt sources. A modular retriever (retriever.retrieve(...)) was used to facilitate potential QA or reranking extensions.

D. LLM Answer Generation

The 'LLMGenerator' component utilized a constrained prompt template, which acted like a 'medical research assistant,' and ensured single token output, such as 'yes,' 'no,' or 'maybe,' along with source reference. Normalization occurred via 'label extractor.'

E. Metric Computation

Performance was measured across:

- Retrieval latency (average per query)

- Generation latency (average per query)
- Embedding latency (index build time)
- Accuracy via PubMedQA and MedMCQA accuracy(predictions, ground_truths)
- F1 score via token overlap (token_f1_score)

F. Visualization

Bar plots were generated for accuracy, F1, and retrieval latency. Axes used short labels (MiniLM-L6-v2, bert-base, BioBERT), rendered at ~150 DPI with annotated values (accuracy/F1 to two decimals, latency to four decimals).

G. System Architecture and Assumptions

The proposed framework of Retrieval-Augmented Generation (RAG) for chatbots in medical contexts consists of a series of steps such as preprocessing, generation of embeddings, retrieval, and answer synthesis, which are modularized together to form a pipeline. It is assumed that PubMedQA and MedMCQA datasets are representative of biomedical queries and that the embedding models are pre-trained enough to generate semantic representations of words, and that FAISS is an efficient vector indexing technique with negligible time lag in retrieval compared to generation and embedding.Reliability and trust of users are considered to be dependent on accuracy and time lag, and interpretability and explainability are considered to be future work.

V. Results and Discussion

MiniLM, BERT, and BioBERT were compared against one another via a comparative analysis and found to have sizeable differences between their: respective accuracy results, semantic robustness, and computational efficiency when used within a RAG-based Medical Chatbot (see Table I for performance metrics of each of the embeddings and their results with a FAISS retrieval infrastructure).

TABLE I: COMPARATIVE EVALUATION OF EMBEDDING MODELS

Model	Accuracy	F1	Doc Embedding Latency (ms)	Query Embedding Latency (ms)	Retrieval Latency (ms)	Generation Latency (s)
MiniLM	0.635	0.525	8.97	0.0083	0.0003	0.8681
BERT	0.490	0.429	32.23	0.0241	0.0003	0.9010
BioBERT	0.670	0.576	27.38	0.0243	0.0002	0.9258

The Table I results showed that BioBERT outperformed all other models in terms of accuracy and F1 measures, which again highlights the importance of pre-training in a biomedical context for semantic understanding. The exceptional efficiency of MiniLM was also evident from the results, which showed a remarkable improvement in reducing the latency of document embedding to 8.97 ms, almost three times faster than BioBERT and BERT. This efficiency of MiniLM makes it highly suitable for use in real-time conversational applications, where efficiency is critical for building trust. Although the accuracy of MiniLM was slightly lower than that of other models, it is evident from the results that MiniLM is a highly efficient model for use in real-time applications.

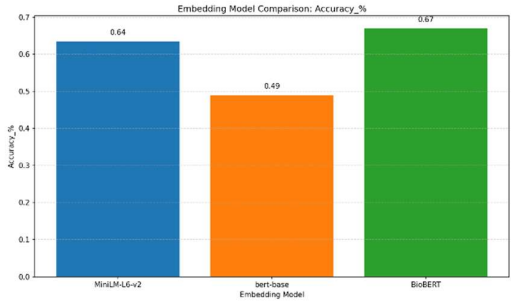

Figure 2

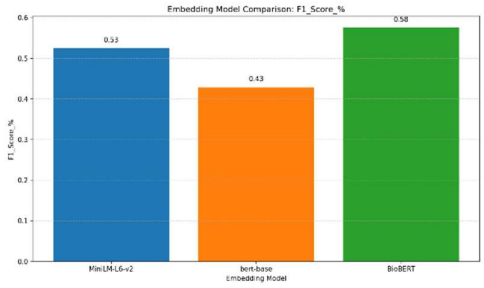

Figure 3

Figure 1. Accuracy comparison of MiniLM, BERT, and BioBERT embeddings on PubMedQA. BioBERT has the highest accuracy of 0.67, while MiniLM provides the best balance of speed and accuracy.

Figure 2. F1 score comparison of MiniLM, BERT, and BioBERT embeddings. BioBERT has the highest F1 score of 0.58. The low latency of retrieval (<0.0003 ms) for all models confirmed the efficiency of FAISS in

building a robust infrastructure for biomedical literature retrieval. The latency of generation was consistent at around 0.9 s for all models, which again confirmed that embedding plays a minor role in generating a sentence. According to the human-centered perspective, these results underscore an inherent trade-off between accuracy and promptness. BioBERT yields the most accurate responses, which are critical for decision support systems, while MiniLM yields the most fluent and prompt responses, which are critical for user-facing systems. This duality mirrors the inherent difficulty of developing medical AI systems, where there is an inherent tension between semantic accuracy and operational efficiency to yield effective systems. The results of this study affirm the literature regarding the effectiveness of domain-specific embeddings for medical AI systems [2], [6] and are consistent with the literature regarding the role of efficiency in determining usability for retrieval-augmented systems [9], [10]. Embedding selection, therefore, is not merely an algorithmic choice but has significant user experience and trust implications for medical AI systems such as chatbots.

VI. Future Work

The future of this work includes evaluating Pinecone, Milvus, multiple pieces of evidence to retrieve and verify the contents, and user-centered research to create a balance between accuracy and timeliness within these medical AI systems. This document provides a comparison of MiniLM, BERT, and BioBERT embeddings in relation to their use within a medical chatbot built with a RAG-based methodology. BioBERT achieved the best performance based on accuracy (i.e., the number of correctly predicted instances) and F1 scores (the combination of precision and recall), thus verifying the effectiveness of using domain-specific data as a foundation for training within the medical usage area. Nevertheless, MiniLM produced significantly greater output speeds than those produced by either BERT or BioBERT; thus, making MiniLM a more suitable model for the development of medical chatbots. All models produced low retrieval times and constant generation times, demonstrating the relationship between the type of embedding and the retrieval efficiency of these medical chatbot systems. Future research will involve the evaluation of additional vector databases, developing a method of combining multiple pieces of evidence to retrieve and verify information, building user-centered trust, and ultimately achieving a balance between accuracy and speed in the realm of medical AI.

VII. Conclusions

This research indicates how embedding choices affect the accuracy-efficiency relationship of Medical Question Answering using RAG-based methods. Semantic accuracy achieved with BioBERT illustrates the benefits of using domain-specific pre-training, and MiniLM has appropriate responsiveness real-time performance. Therefore, MiniLM has high potential for use within patient interaction systems. This research provides direction for future research about scalable, trustworthy medical chatbot systems while balancing precision vs cutting-edge responsiveness through the exploration of either multi-evidence retrieval or diverse vector databases, and user-centered trust analyses that all support/trust in Healthcare AI.

References

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T. and Riedel, S., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, pp.9459-9474.

[2] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), pp.1234-1240.

[3] Peng, Y., Yan, S. and Lu, Z., 2019, August. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In Proceedings of the 18th BioNLP workshop and shared task (pp. 58-65).

[4] Jin, Q., Dhingra, B., Liu, Z., Cohen, W. and Lu, X., 2019, November. Pubmedqa: A dataset for biomedical research question answering. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 2567-2577).

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. Advances in neural information processing systems, 33, pp.1877-1901.

[6] Beltagy, I., Lo, K. and Cohan, A., 2019, November. SciBERT: A pretrained language model for scientific text. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 3615-3620).

[7] Guu, K., Lee, K., Tung, Z., Pasupat, P. and Chang, M., 2020, November. Retrieval augmented language model pre-training. In International conference on machine learning (pp. 3929-3938). PMLR.

[8] Khattab, O. and Zaharia, M., 2020, July. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval (pp. 39-48).
[9] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W.T., 2020, November. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 6769-6781).
[10] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A. and Gurevych, I., 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. arXiv preprint arXiv:2104.08663.
[11] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019, June. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).
[12] R. Nogueira and K. Cho, "Passage Re-ranking with BERT," arXiv preprint arXiv:1901.04085, 2019.
[13] Y. Zhang et al., "MedBioRAG: Semantic Search and Retrieval-Augmented Generation with Large Language Models for Medical Question Answering," arXiv preprint arXiv:2512.10996, 2025.
[14] A. Garg, S. Sharma, and R. Gupta, "Biomedical Literature Question Answering System Using Retrieval-Augmented Generation," arXiv preprint arXiv:2509.05505, 2025.
[15] Alamri, A.A., Borik, R.M., El-Wahab, A.H.A., Mohamed, H.M., Ismail, K.S., El-Aassar, M.R., Al-Dies, A.A.M. and El-Agrody, A.M., 2025. Synthesis of Schiff bases based on Chitosan, thermal stability and evaluation of antimicrobial and antitumor activities. Scientific Reports, 15(1), p.892.
[16] Ryan, J. et al. (2026) Self-medrag: A self-reflective hybrid retrieval-augmented generation framework for reliable medical question answering, arXiv.org. Available at: https://arxiv.org/abs/2601.04531 (Accessed: 31 March 2026).
[17] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S. and Payne, P., 2023. Large language models encode clinical knowledge. Nature, 620(7972), pp.172-180.
[18] M. Chen et al., "Evaluating Large Language Models Trained on Medical Question Answering Datasets," Nature Digital Medicine, vol. 6, pp. 1–9, 2023.
[19] Rajapaksha, S., Kalutarage, H., Al-Kadri, M.O., Petrovski, A., Madzudzo, G. and Cheah, M., 2023. Ai-based intrusion detection systems for in-vehicle networks: A survey. ACM Computing Surveys, 55(11), pp.1-40.
[20] S. Yang et al., "Retrieval-Augmented Generation in Healthcare: Methods and Applications," IEEE Access, vol. 13, pp. 21534–21548, 2025.