Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

Conversational Agent for Medical Question-Answering Using RAG and LLM

La Ode Muhammad Yudhy Prayitno[1]*, Annisa Nurfadilah [2], Septiyani Bayu Saudi [3], Widya Dwi Tsunami [4], Adha Mashur Sajiah[5]

[1,2,3,4,5] Teknik Informatika, Universitas Halu Oleo
yudhyprayitno567@gmail.com[1], anisanurfadilah2406@gmail.com[2], septiyanibayusaudi@gmail.com[3], widyadwitsunami04@gmail.com[4], adha.m.sajiah@uho.ac.id[5]*

Abstract

This study analyzes the application of the RAG concept alongside an LLM in the context of PubMed QA data to augment question-answering capabilities in the medical context. For answering questions relevant to private healthcare institutions, the Mistral 7B model was utilized. To limit hallucinations, an embedding model was used for document indexing, ensuring that the LLM answers based on the provided context information. The analysis was conducted using five embedding models, two of which are specialized medical models, PubMedBERT-base and BioLORD-2023, as well as three general models, GIST-large-Embedding-v0, blade-embed-kd, and all-MiniLM-L6-v2. As the results showed, general models performed better than domain specific models, especially GIST-large-Embedding-v0 and b1ade-embed-kd, which underscores the dominance of general-purpose training datasets in terms of fundamental semantic retrieval, even in medical domains. The outcome of this research study demonstrates that applying RAG and LLM locally can safeguard privacy while still responding to medical queries with appropriate precision, thus establishing a foundation for a dependable medical question-answering system.

Keywords: *Embedding Models; Large Language Model; Medical Question-Answering; PubMed; Retrieval-Augmented Generation*

1. Introduction

Conversational agents, also called chatbots, have become tremendous since the first chatbot ELIZA appeared in the early 1960s [4][12]. The concept of a chatbot was initiated in the Turing test in 1950, with various rapid developments in data and technology over five decades, and modern chatbots now build upon LLMs [6]. Large Language Models are advances of transformer-based large language models (LMs) that predict the probability of a sequence of tokens occurring based on content knowledge from large datasets [1], [3], [5]. For example, ChatGPT, developed by OpenAI, has quickly gained popularity and significantly impacted various fields, including education and research [2], [7].

But with the power of generation context, users have significant privacy and data protection concerns when using this system [8]. Also, in electronic healthcare there are big concerns about security and privacy risks for personal data. The major concern is each country having varying standards for the security and privacy of medical data [9]. Third-party LLMs such as OpenAI's ChatGPT raise significant privacy concerns in the medical sector, as sensitive patient data must be sent externally, where information about the model training process or the underlying data is often limited [10].
The solution is training with locally deployable LLMs, where the model is more secure as it can be trained and operated on-premise, keeping medical data within the control of the institution. This approach marks a significant step forward in privacy-preserving technologies for healthcare, allowing us to leverage the immense potential of digital health records for research and clinical advancement without sacrificing patient privacy [11]. Nevertheless, LLMs also have disadvantages like answering questions with out-of-scope information or generating plausible yet nonfactual content, which is called LLM hallucination [13].

The existing solution is Retrieval-Augmented Generation (RAG), which is an approach to enhance Large Language Models (LLMs) by combining their parametric knowledge with external knowledge sources to address issues like hallucinations and outdated information, by dynamically augmenting the model's capabilities with up-to-date medical data [14]. To ensure high-quality RAG, the embedding model is a vital step to make retrieval have high similarity and quality system performance [15]. In this study, the implementation of a RAG pipeline for question-answering with medical data using PubMed QA [16] is proposed, employing a variety of embedding models for semantic medical text [17]. For generating answers, Mistral 7B was utilized as the Large Language Model, which has better performance for medical tasks [18].

2. Research Methodology

This section explains the research methods applied in this study for implementing a RAG pipeline in a medical question-answering system with Mistral 7B as the local large language model. The focus of this research is to identify the best embedding models for semantic retrieval in the medical domain using the PubMed QA dataset, with semantic answer similarity evaluator (SAS), mean reciprocal rank (MRR), and Faithfulness as evaluation metrics. The following is the flow of the research methodology, which can be seen in Fig. 1 below.

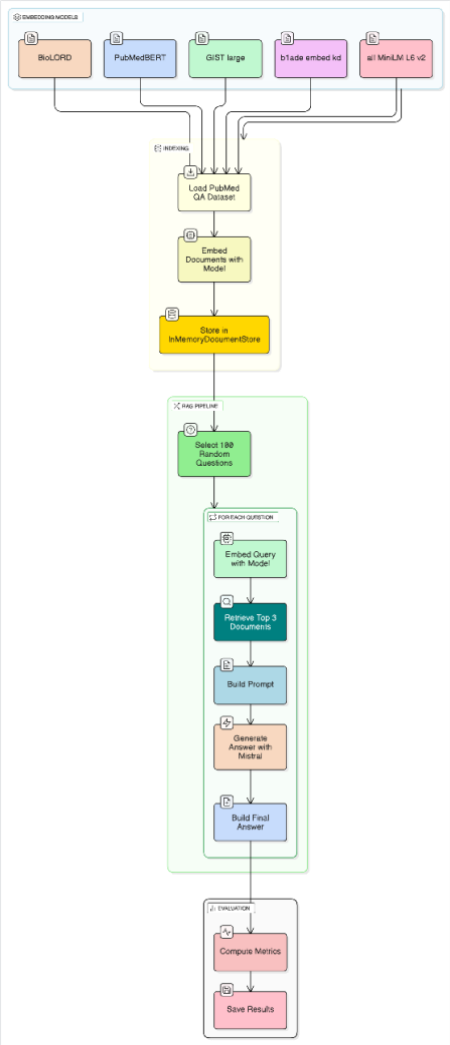

Fig. 1: Research workflow

2.1. Data Collection

In this study, the dataset used is the PubMed QA dataset containing medical question-answer pairs obtained from PubMed abstracts. This dataset is of high quality because the medical questions and their corresponding answers have been verified by professionals in the field. This dataset was selected because it captures real-life medical situations and contains vast biomedical information that is useful for testing RAG systems in the healthcare field.

2.2. Document Indexing

Document indexing is the first stage where the PubMed QA dataset is processed and made ready for the retrieval system. In this phase, all documents in the dataset are embedded by five different models: PubMedBERT-base, GIST-large-Embedding-v0, BioLORD-2023, b1ade-embed-kd, and all-MiniLM-L6-v2. Each of these models creates vector representations of the medical documents, which are later stored in InMemoryDocumentStore for quick retrieval.

2.3. RAG Pipeline Implementation

The RAG pipeline implementation consists of specific pivotal stages for dealing with questions in the medical domain. To begin with, 100 randomly selected questions were set aside from the dataset for assessment purposes. For every question, the system embeds the query using the same model used for embedding the documents, which guarantees uniformity within the system. After doing this, the system retrieves the top 3 most relevant documents using a semantic similarity scoring system. Then, a ChatPromptBuilder creates a prompt using the retrieved documents and the question, followed by generating an answer with OllamaChatGenerator using the Mistral model. Finally, the AnswerBuilder processes the final answer.

2.4. Model Evaluation

In the model evaluation step, the system performance is analyzed using multiple metrics for complete assessment:

1. Mean Reciprocal Rank (MRR): Evaluates the ranking by retrieving relevant documents and assessing each document's rank (position).
2. Semantic Answer Similarity (SAS): Employing advanced modern metrics, SAS measures the provided answer's semantic proximity to the actual answer, determining how closely they correspond.
3. Faithfulness: Evaluates if the provided answers are consistent and faithful to the retrieved documents without hallucinated information.

Results are saved to a CSV file and the process is repeated for each embedding model to find the best embedding models.

3. Result and Discussion

3.1. Data Collection

The data used consists of text collected from Hugging Face PubMedQA_instruction with a total of 273k rows, separated into 272k rows as train and 1k rows as test. The original data has 4 columns: instruction refers to the question, context as the document reference which contains abstracts from PubMed articles, response is the correct answer based on context, and category is designated for the question-answering that is already done with closed_qa. The following Fig. 2 shows the data collected from PubMedQA_instruction on HuggingFace.

instruction string	context string	response string	category string
Are group 2 innate lymphoid cells (ILC2s) increased in chronic rhinosinusitis…	Chronic rhinosinusitis (CRS) is a heterogeneous disease with an uncertain…	As ILC2s are elevated in patients with CRSwNP, they may drive nasal polyp…	closed_qa
Does vagus nerve contribute to the development of steatohepatitis and…	Phosphatidylethanolamine N-methyltransferase (PEMT), a liver…	Neuronal signals via the hepatic vagus nerve contribute to the development of…	closed_qa
Does psammaplin A induce Sirtuin 1-dependent autophagic cell death in…	Psammaplin A (PsA) is a natural product isolated from marine sponges, which has…	PsA significantly inhibited MCF-7/adr cells proliferation in a concentration…	closed_qa
Is methylation of the FGFR2 gene associated with high birth weight centil…	This study examined links between DNA methylation and birth weight centile…	We identified a novel biologically plausible candidate (FGFR2) for with…	closed_qa
Do tumor-infiltrating immune cell profiles and their change after…	Tumor microenvironment immunity is associated with breast cancer outcome. A…	Breast cancer immune cell subpopulation profiles, determined by…	closed_qa

Fig. 2: Sample rows from the PubMedQA_instruction dataset

In this process, related columns were selected to ensure that the model can generate answers based on the context. This research down-sampled the first 10,000 rows from the train sample of the PubMedQA instruction dataset, with selected columns: instruction as all questions to test the model later, context as all documents for retrieval to embedding models, and response as the ground truth answer to evaluate similarity from the generated LLM answer.

3.2. Document Indexing

After collecting the PubMedQA_instruction as described in the previous section, the next step is the document indexing process. The 10,000 documents (from the context column) were processed to create vector representations using five embedding models for comparison, namely PubMedBERT-base [20], GIST-large-Embedding-v0 [21], BioLORD-2023 [22], blade-embed-kd [23], and all-MiniLM-L6-v2 [24]. The selection of these models is based on their prevalence in the field of biomedicine and general NLP, and their performance to allow a comprehensive comparison of the quality of retrieval and answer generation results in the medical domain. Each embedding model has different characteristics that may affect the quality of semantic search. For example, PubMedBERT-base and BioLORD-2023 are specifically trained on biomedical corpora and therefore better understand complex medical contexts. Meanwhile, GIST-large-Embedding-v0 is fine-tuned for high-performance text retrieval tasks across multiple domains. The b1ade-embed-kd model is a version trained from large data by learning from other larger models (distillation), while all-MiniLM-L6-v2 is designed for speed, with a smaller model size and a limited context window. The following Table 1 shows the specifications of each embedding model:

Table 1: Specifications of embedding models used for document indexing

Embedding Model	Window Size (Tokens)	Parameters (Millions)	Dimensions
PubMedBERT-base	512	110	768
GIST-large-Embedding-v0	512	335	1024
BioLORD-2023	512	109	768
b1ade-embed-kd	512	335	1024
all-MiniLM-L6-v2	256	22.7	384

All embedded documents are then stored as embeddings along with their respective documents in the vector database. A vector database is a type of database that stores data as high-dimensional vectors, which are mathematical representations of features or attributes [19]. SentenceTransformDocumentEmbedder from Haystack was used to convert text into dense vectors, which were then stored in InMemoryDocumentStore using DocumentWriter.

3.3. RAG Pipeline Implementations

After the document indexing process is complete and the entire medical context is successfully converted into vector form, the next step is the implementation of the RAG pipeline. The pipeline has a few main steps:

1. Query Embedding: Each question was embedded using the same model as the documents (e.g., PubMedBERT for the PubMedBERT store) to ensure consistency.
2. Document Retrieval: InMemoryEmbeddingRetriever was used to retrieve the top-3 documents based on cosine similarity between query and document embeddings.
3. Prompt Building: ChatPromptBuilder creates a prompt with the retrieved documents and question to guide the LLM. In this case, the template for the LLM makes it answer like this:

```
# Define the prompt template
template = [
    ChatMessage.from_user(
        """
        You are a medical expert answering questions based on the provided context. Use only the context
        to answer the question accurately.

        Context:
        {% for document in documents %}
            {{ document.content }}
        {% endfor %}

        Question: {{question}}
        Answer:
        """
    )
]
```

Fig. 3: Prompt template used to guide the LLM based on retrieved document context

   The LLM was implemented to answer based on the context provided from document retrieval, where the LLM only answers the question using the top 3 documents with the highest similarity. This ensures that the LLM does not hallucinate when answering the question, as the LLM is provided with the relevant document context.
4. Answer Generation: Mistral 7B, run locally with Ollama, was used to generate an answer from the prompt.
5. Final Answer: AnswerBuilder cleans up the generated answer for evaluation.

The following Fig. 4 illustrates the RAG pipeline flow developed in this study.

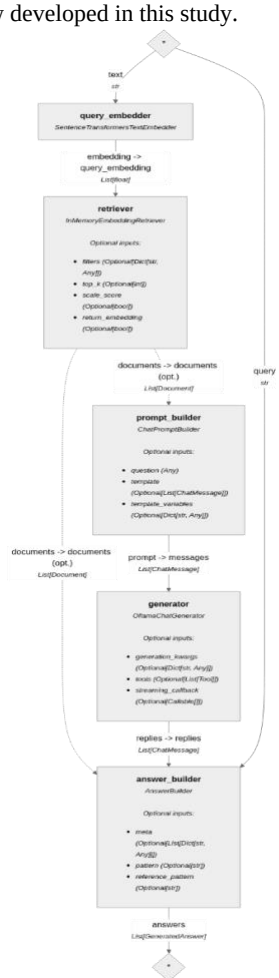

Fig. 4: RAG pipeline flow

3.4. Model Evaluation

To evaluate system performance, 100 questions were randomly selected from the test set using random.seed(42) to ensure reproducibility. For each embedding model, the RAG pipeline generated answers based on the retrieved documents, which were then evaluated using three metrics: Mean Reciprocal Rank (MRR) to assess document retrieval quality, Semantic Answer Similarity (SAS) to

measure alignment with ground-truth answers, and FaithfulnessEvaluator to verify factual consistency with retrieved contexts. Below is an example of answers generated by different embedding models for the question:
"Do high levels of procalcitonin in the early phase after pediatric liver transplantation indicate poor postoperative outcome?"

Table 2: Example of generated answers for a sample question using different embedding models

Embedding Model	Generated Answer
PubMedBERT-base	Yes, according to the provided context, high levels of procalcitonin in the early phase after pediatric liver transplantation do indicate a poor postoperative outcome. Patients with high PCT levels on postoperative day 2 had higher International Normalized Ratio values on POD 5, suffered more often from primary graft non-function, had a longer stay in the pediatric intensive care unit and on mechanical ventilation. However, there was no correlation between PCT elevation and systemic infection.
GIST-large-Embedding-v0	Yes, according to the provided context, high levels of procalcitonin (PCT) on postoperative day 2 in pediatric liver transplant recipients are associated with a higher International Normalized Ratio value on postoperative day 5, more instances of primary graft non-function, longer stays in the pediatric intensive care unit, and longer duration on mechanical ventilation. However, it is important to note that PCT levels were not correlated with systemic infection.
BioLORD-2023	Yes, according to the provided context, high levels of procalcitonin (PCT) in the early phase after pediatric liver transplantation are associated with a poorer postoperative outcome. Specifically, patients with high PCT levels on postoperative day 2 had higher International Normalized Ratio values on postoperative day 5, suffered more often from primary graft non-function, had a longer stay in the pediatric intensive care unit and on mechanical ventilation. However, it is important to note that there was no correlation between PCT elevation and systemic infection.
b1ade-embed-kd	Yes, based on the provided context, patients with high PCT levels on postoperative day 2 had a higher incidence of primary graft non-function, longer stay in the pediatric intensive care unit, and a longer duration on mechanical ventilation. They also had higher International Normalized Ratio values on postoperative day 5. Therefore, high PCT levels in the early phase after pediatric liver transplantation appear to indicate a poor postoperative outcome.
all-MiniLM-L6-v2	Yes, based on the provided context, high levels of procalcitonin (PCT) in the early phase after pediatric liver transplantation are associated with a higher International Normalized Ratio (INR) value on postoperative day 5, a longer stay in the pediatric intensive care unit ($p<0.01$), and a longer duration of mechanical ventilation ($p=0.001$). Additionally, patients with high PCT levels had a higher incidence of primary graft non-function ($p<0.05$). These findings suggest that high levels of PCT in the early phase after pediatric liver transplantation may indicate a poor postoperative outcome. However, it is important to note that there was no correlation between PCT elevation and systemic infection in this study.

Summarizes the average scores across 100 questions for each embedding model, which can be seen in Table. 3 below.

Table 3: Average scores across 100 questions for each embedding model

Embedding Model	MRR	SAS	Faithfulness
PubMedBERT-base	0.985	0.718	0.949
GIST-large-Embedding-v0	**1.000**	0.814	**0.950**
BioLORD-2023	0.910	0.676	0.939
b1ade-embed-kd	0.866	**0.855**	0.916
all-MiniLM-L6-v2	0.975	0.715	0.932

As observed in Table 3, the performance of each model differs when tested using the three evaluation metrics (MRR, SAS, Faithfulness). The best performance is held by GIST-large-Embedding-v0 with an MRR score of 1.000 and Faithfulness of 0.950, while the highest score on SAS is held by the b1ade-embed-kd model. It can be observed that, in the case of question-answering using the RAG technique with the LLM model as the answer generator, where answers are obtained from retrieval results of the top 3 documents using the embedding model for document indexing, the best performance is achieved by general embedding models, while embedding models trained with medical data generally underperform. This means that general models are superior to biomedical models due to their training on broader and more diverse data [25]. General models such as by GIST-large-Embedding-v0, b1ade-embed-kd, and all-MiniLM-L6-v2, especially those trained on diverse data, are more resilient to input variations and provide more accurate search results in medical contexts. This suggests that domain-specific models (PubMedBERT-base, BioLORD-2023) are not necessarily better for medical tasks, where diversity of training data is more important for achieving better performance in medical search.

4. Conclusion

In this study, the RAG concept with LLM was successfully applied to generate answers in question-answering using PubMed QA data to address medical-related questions. The LLM with the Mistral 7B model was utilized locally to generate answers, focusing on private use for healthcare institutions. In addition, the embedding model is used to index the documents so that the LLM model can only answer based on the available context data to avoid hallucinatory answers. In the application, 5 different embedding models are used: PubMedBERT-base and BioLORD-2023 models are models trained on medical field corpora, and GIST-large-Embedding-v0, b1ade-embed-kd, and all-MiniLM-L6-v2 models are models trained on a variety of different domains. It was found that general embedding models, specifically GIST-large-Embedding-v0 and b1ade-embed-kd, outperformed domain-specific models such as PubMedBERT and BioLORD-2023 in terms of MRR, SAS, and Faithfulness metrics. GIST-large-Embedding-v0 demonstrated superior retrieval performance with perfect MRR scores (1.000) and high faithfulness (0.950), while b1ade-embed-kd excelled in semantic similarity (0.855), both significantly outperforming specialized medical models. This suggests that broader training data improves semantic retrieval and answer generation in a medical context. Leveraging RAG and LLM locally can minimize privacy problems while still maintaining high-quality responses, which can be a solution for a safe and reliable medical question-answering system.

References

[1] D. Erlansyah, A. Mukminin, D. Julian, E. S. Negara, F. Aditya, and R. Syaputra, "Large language model (LLM) comparison between GPT-3 and PaLM-2 to produce Indonesian cultural content", *EEJET*, vol. 4, no. 2 (130), pp. 19–29, Aug. 2024, doi: https://doi.org/10.15587/1729-4061.2024.309972.

[2] Miah, et al., "ChatGPT in Research and Education: Exploring Benefits and Threats," arXiv (Cornell University), Nov. 2024, doi: https://doi.org/10.48550/arxiv.2411.02816.

[3] S. Minaee et al., "Large Language Models: A Survey," arXiv (Cornell University), Feb. 2024, doi: https://doi.org/10.48550/arxiv.2402.06196.

[4] J. Shrager, "ELIZA Reinterpreted: The world's first chatbot was not intended as a chatbot at all," arXiv.org, Jun. 25, 2024. https://arxiv.org/abs/2406.17650.

[5] T. Xiao and J. Zhu, "Foundations of Large Language Models," arXiv (Cornell University), Jan. 2025, doi: https://doi.org/10.48550/arxiv.2501.09223.

[6] J. Xue, Y. Wang, C. Wei, X. Liu, J. Woo, and C.-C. Jay Kuo, "Bias and Fairness in Chatbots: An Overview," arXiv (Cornell University), Sep. 2023, doi: https://doi.org/10.48550/arxiv.2309.08836.

[7] J.-J. Zhu, J. Jiang, M. Yang, and Z. J. Ren, "ChatGPT and Environmental Research," Environmental Science & Technology, vol. 57, no. 46, Mar. 2023, doi: https://doi.org/10.1021/acs.est.3c01818.

[8] G. Sebastian, "Privacy and Data Protection in ChatGPT and Other AI Chatbots: Strategies for Securing User Information," International Journal of Security and Privacy in Pervasive Computing, vol. 15, no. 1, Jan. 2023, doi: https://doi.org/10.2139/ssrn.4454761.

[9] V. Mishra, K. Gupta, D. Saxena, and Ashutosh Kumar Singh, "A Global Medical Data Security and Privacy Preserving Standards Identification Framework for Electronic Healthcare Consumers," IEEE Transactions on Consumer Electronics, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/tce.2024.3373912.

[10] Cody, A. Mullen, S. Armstrong, C. Hickey, and J. Talbert, "Local Large Language Models for Complex Structured Medical Tasks," arXiv (Cornell University), Jan. 2023, doi: https://doi.org/10.48550/arxiv.2308.01727.

[11] I. C. Wiest, M.-E. Lessmann, F. Wolf, D. Ferber, and J. N. Kather, "Anonymizing medical documents with local, privacy preserving large language models: The LLM-Anonymizer," Jun. 13, 2024. https://www.researchgate.net/publication/381417636_Anonymizing_medical_documents_with_local_privacy_preserving_large_language_models_The_LLM-Anonymizer

[12] R. Sutcliffe, "A Survey of Personality, Persona, and Profile in Conversational Agents and Chatbots," arXiv (Cornell University), Jan. 2024, doi: https://doi.org/10.48550/arxiv.2401.00609.

[13] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," ACM transactions on office information systems, vol. 43, no. 2, Nov. 2024, doi: https://doi.org/10.1145/3703155.

[14] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv.org, Dec. 18, 2023. https://arxiv.org/abs/2312.10997.

[15] L. Caspari, D. K. Ghosh, S. Zerhoudi, J. Mitrovic, and M. Granitzer, "Beyond Benchmarks: Evaluating Embedding Model Similarity for Retrieval Augmented Generation Systems," arXiv (Cornell University), Jul. 2024, doi: https://doi.org/10.48550/arxiv.2407.08275.

[16] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "PubMedQA: A Dataset for Biomedical Research Question Answering," arXiv.org, Sep. 13, 2019. https://arxiv.org/abs/1909.06146.

[17] S. Soffer et al., "A Scalable Framework for Benchmarking Embedding Models for Semantic Medical Tasks," Aug. 2024, doi: https://doi.org/10.1101/2024.08.14.24312010.

[18] Diash Firdaus, Idi Sumardi, and Yuni Kulsum, "Integrating Retrieval-Augmented Generation with Large Language Model Mistral 7b for Indonesian Medical Herb," JISKA (Jurnal Informatika Sunan Kalijaga), vol. 9, no. 3, pp. 230–243, Sep. 2024, doi: https://doi.org/10.14421/jiska.2024.9.3.230-243.

[19] Y. Han, C. Liu, and P. Wang, "A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge," arXiv.org, Oct. 18, 2023. https://arxiv.org/abs/2310.11703.

[20] Y. Gu et al., "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," ACM Transactions on Computing for Healthcare, vol. 3, no. 1, pp. 1–23, Jan. 2022, doi: https://doi.org/10.1145/3458754.

[21] Solatorio, Aivin V, "GISTEmbed: Guided In-sample Selection of Training Negatives for Text Embedding Fine-tuning," arXiv (Cornell University), Feb. 2024, doi: https://doi.org/10.48550/arxiv.2402.16829.

[22] F. Remy, K. Demuynck, and T. Demeester, "BioLORD-2023: Semantic Textual Representations Fusing LLM and Clinical Knowledge Graph Insights," arXiv (Cornell University), Jan. 2023, doi: https://doi.org/10.48550/arxiv.2311.16075.

[23] "w601sxs/b1ade-embed-kd · Hugging Face," Huggingface.co, 2024. https://huggingface.co/w601sxs/b1ade-embed-kd (accessed May 27, 2025).

[24] C. Yin and Z. Zhang, "A Study of Sentence Similarity Based on the All-minilm-l6-v2 Model With 'Same Semantics, Different Structure' After Fine Tuning," Advances in computer science research, pp. 677–684, Jan. 2024, doi: https://doi.org/10.2991/978-94-6463-540-9_69.

[25] J.-B. Excoffier, T. Roehr, A. Figueroa, M. Papaaioannou, K. Bressem, and M. Ortala, "Generalist embedding models are better at short-context clinical semantic search than specialized embedding models," arXiv (Cornell University), Jan. 2024, doi: https://doi.org/10.48550/arxiv.2401.01943.