Enhancing retrieval performance of embedding models via fine-tuning on synthetic data in RAG chatbot for Vietnamese military science domain

Nguyen Xuan Bac, Luu Van Sang, Nguyen Duc Vuong,
Luong Quoc Le, Dang Duc Thinh*

Institute of Information Technology, Academy of Military Science and Technology, 17 Hoang Sam, Cau Giay, Hanoi, Vietnam.
*Corresponding author: dangducthinh195@gmail.com
Received: 09 Sep. 2024; Revised 08 Nov. 2024; Accepted 12 Nov. 2024; Published 25 Nov. 2024.
DOI: https://doi.org/10.54939/1859-1043.j.mst.99.2024.109-118

ABSTRACT

Retrieval-Augmented Generation (RAG) is a technology that combines information retrieval with large language models, enabling chatbots to provide accurate answers by querying relevant documents from a data repository before generating responses. While RAG chatbot has demonstrated effectiveness in many applications, there are still limitations in specialized Vietnamese data domains, particularly in the military science field. To address this challenge, this paper proposes a framework for fine-tuning embedding models using synthetic datasets generated by ChatGPT to enhance retrieval performance in a Q&A application focused on the history of the Institute of Information Technology (IoIT). Our approach evaluates 11 popular embedding models and shows a significant average improvement of 18.15% in the MAP@K metric. The resulting IoIT history Q&A chatbot, developed with fine-tuned embedding models and the Vietnamese language model Vistral-7B, outperforms chatbots utilizing OpenAI's embedding models and ChatGPT. These findings highlight the potential of RAG chatbot technology for advancing information retrieval applications in specialized fields like military science.

Keywords: Retrieval-augmented generation; Fine-tuning; Synthetic data; Large language model; Chatbot.

1. INTRODUCTION

Conversational agents powered by large language models (LLMs), particularly chatbots, have significantly improved natural language understanding and generation capabilities. However, these chatbots exhibit several limitations that affect their performance, particularly when applied to real-world and specialized domains. One major limitation of LLMs is the outdated nature of their knowledge [1, 2]. Since LLMs are trained on static datasets up to a specific cutoff date, they are unable to incorporate new information that emerges post-training. As a result, LLMs may generate responses based on obsolete or incorrect information, making them less reliable for dynamic or evolving fields. Another prominent limitation is the lack of specialized knowledge or in-depth expertise [1]. LLMs are often trained on broad, generalized datasets that prioritize coverage across diverse topics. This results in a shallow understanding of niche domains, such as military science or historical data, where the accurate use of terminology and concepts is crucial. For instance, chatbots built on LLMs often struggle with retrieving precise answers when asked about highly specialized or technical subjects, such as the history and development of institutions like the Institute of Information Technology (IoIT). This lack of depth hinders the ability of LLM chatbots to provide authoritative answers for tasks that demand expertise.

In contrast, Retrieval-Augmented Generation (RAG) technology enhances chatbot performance by combining information retrieval with LLMs [3]. RAG systems address

the limitations of LLMs by querying relevant documents from a data repository before generating responses, ensuring that the chatbot has access to up-to-date and domain-specific information. This capability enables RAG-based chatbots to deliver more accurate and reliable answers, particularly in specialized fields like military science, where precision and contextual understanding are paramount. Despite the significant progress made with RAG-based systems in general domains, there remain persistent challenges when applied to more specialized fields and low-resource languages, such as military science and Vietnamese data domains.

The Vietnamese language, with its complex structures and specialized vocabulary in fields such as military science, presents challenges for effective retrieval systems. General embedding models like PhoBERT [4] and Bkai-vi-bi-encoder [5] have been developed for Vietnamese language tasks, but they often struggle with domain-specific terminology. Models like Vi-LegalText-SBERT [6] for the legal field and ViPubmedDeBERTa [7] for the medical domain have demonstrated the value of fine-tuning in capturing specialized language. These examples highlight the necessity of fine-tuning models on domain-specific data to enhance retrieval performance, particularly in fields requiring specialized knowledge.

This paper focuses on enhancing the retrieval performance of embedding models within RAG chatbots, specifically tailored to the history of the IoIT. By leveraging synthetic data generated by ChatGPT, we propose a framework to fine-tune opensource embedding models to improve their performance in retrieving historical and military science-related data. The fine-tuning process enables these models to better understand and retrieve information within the context of the Vietnamese language and specialized terms. Experimental results show that fine-tuning open-source embedding models not only enhances performance but also has the potential to outperform closed-source and paid solutions like OpenAI's embedding models [8]. By incorporating the fine-tuned models into a RAG-based Q&A chatbot, we demonstrate the superiority of our approach, particularly when paired with the Vietnamese large language model Vistral-7B [9], over existing solutions like OpenAI’s embedding models and ChatGPT. This finding unlocks numerous chatbot applications in highly confidential domains such as military and finance, where data privacy is paramount.

The contributions of this paper are threefold: (1) we introduce a novel approach for fine-tuning embedding models on synthetic data in specialized Vietnamese domains, (2) we validate our approach by significantly improving retrieval performance in a real-world Q&A application about IoIT history, and (3) we demonstrate the potential of RAG technology in building advanced information retrieval systems for complex, specialized domains, with a focus on military science.

2. RELATED WORKS

Embeddings are dense vector representations of words or phrases, designed to capture semantic meaning and relationships in textual data. They are a cornerstone of modern natural language processing, providing a way to represent words in a continuous vector space where similar words are positioned closer together. Popular embedding models include Word2Vec [10], GloVe [11], and transformer-based models such as BERT and its variants [12].

Recent advancements in embedding models have employed a two-step process: initial training on large-scale text corpora, followed by fine-tuning on large, weakly-supervised question-context pairs using contrastive loss [3]. This approach results in what is referred to as pre-trained embedding models (PTE models). Building a PTE model involves training on extensive text data to develop a broad understanding of language and then fine-tuning it on specific tasks to adapt the model to particular contexts or domains. This method enhances the model's ability to handle specialized queries and context-sensitive tasks effectively.

For example, popular open-source PTE models include BGE Embedding [13], ColBERT for English [14]; Multilingual E5 designed to handle multiple languages [15]. Recently, OpenAI introduced new text-embedding models, including text-embedding-3-small and text-embedding-3-large, which have demonstrated strong multilingual performance. For Vietnamese, general embedding models like PhoBERT [4] and Bkai-vi-bi-encoder [5] have been developed to address the language's unique structures and vocabulary.

Despite their advancements, these models may still struggle with domain-specific terminology. Fine-tuning on domain-specific Vietnamese data, including historical or military contexts, helps the model capture relevant nuances and improve its performance in these specialized areas.

3. DATA

Data used in this paper is the historical document titled *"History of the Institute of Information Technology (1974–2019)"*. This 192-page book, written in Vietnamese, offers a detailed account of the establishment, development, and achievements of the IoIT over 45 years.

The document traces the IoIT's origins to 1974, during the final stages of the Vietnam War, when it was first established as the Mathematics-Computing Department under the Military Engineering Institute. Significant milestones include the installation of the Soviet-made Minsk-32 computer in Hanoi in 1974, marking the birth of Vietnam's military computing sector. The book documents the IoIT's organizational transitions and contributions to national defense, scientific research, and military technology.

The document uniquely combines military domain knowledge with scientific advancements, offering a challenging dataset for any retrieval-augmented generation (RAG) chatbot application. The highly specialized terminology and content in both military and scientific contexts make it essential to fine-tune pre-trained embedding models for improved retrieval performance. By addressing these complex fields in the Vietnamese language, we aim to demonstrate the effectiveness of fine-tuning to enhance the chatbot's ability to retrieve accurate and contextually appropriate information within these specialized domains.

To test retrieval performance and the effectiveness of our RAG Chatbot, we constructed a test dataset consisting of 67 questions, along with the ground truth answers and the corresponding ground truth contexts. This dataset was curated by experts to ensure a high level of accuracy and relevance.

4. METHOD

4.1. Overview of the proposed fine-tuning framework

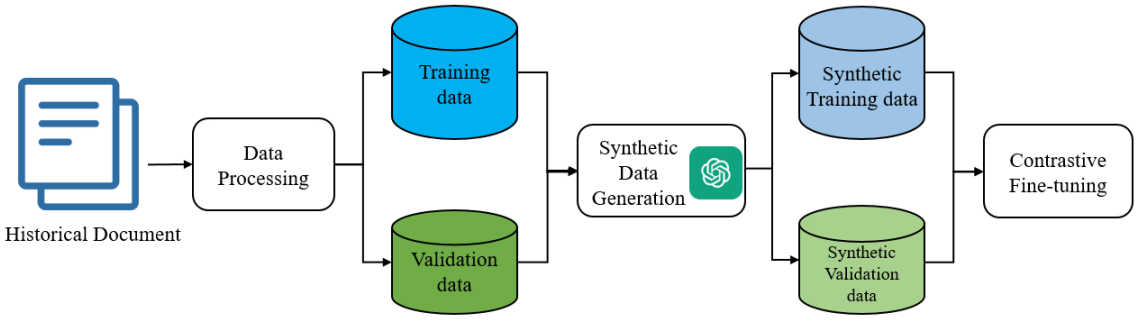

Figure 1. *Overview of the proposed fine-tuning framework.*

In this paper, we introduce a fine-tuning framework using contrastive learning on synthetic data generated by ChatGPT-3.5. The three-phase framework is illustrated in figure 1: (1) Data Processing Phase, (2) Synthetic Data Generation Phase, and (3) Contrastive Fine-Tuning Phase.

4.2. Data processing

In data processing phase, we processed the document titled "History of the Institute of Information Technology (1974–2019)" in two steps:

- ***Step 1*****:** The document was chunked into contexts, where each context represents a full paragraph. This method ensures that each context contains complete semantic information, as opposed to chunking by a fixed number of tokens or sentences. The 192-page document resulted in 515 distinct contexts.

- ***Step 2*****:** These contexts were then split into two datasets using a 90:10 ratio. This resulted in a training dataset containing 463 contexts and a validation dataset containing 52 contexts.

Table 1. *Description of datasets.*

Dataset	**Raw (contexts)**	**After generation (qc pairs)**
Training dataset	463	2315
Validation dataset	52	260
Total	*515*	*2575*

4.3. Synthetic data generation

The embedding model plays a crucial role in retrieving the most relevant context that contains the necessary information to answer a user's question. Therefore, each training sample requires a pair of questions and its corresponding relevant context to improve retrieval performance. Manually generating these question-context pairs would be highly time-consuming and labor-intensive. To address this, we leverage the capabilities of LLMs, specifically ChatGPT-3.5, which is recognized as an efficient training data generator and be integrated into popular data generation frameworks such as LlamaIndex [16] or RAGAS [17] to automate the process.

We utilized the OpenAI API with the following prompt to generate five questions for each context, based on the size of the raw data:

"You are a teacher. Your task is to setup 5 questions for the upcoming quiz/test based

on the provided context information, without relying on prior knowledge. The questions should be diverse and cover the entire context. Each question should begin with "Question:".

Context: {context_str}"

As shown in table 1, this approach yielded 2,315 question-context (qc) pairs for training dataset and 260 qc pairs for validation dataset after processing.

4.4. Contrastive fine-tuning

In this phase, we apply a contrastive learning approach using Multiple Negatives Ranking (MNR) Loss [18], which is a widely adopted contrastive loss technique for fine-tuning embedding models. The goal of MNR Loss is to enhance the model's ability to retrieve relevant contexts by optimizing the cosine similarity between question-context pairs, particularly focusing on maximizing the similarity for positive pairs (correctly matched question-context pairs) while minimizing it for negative pairs (mismatched pairs).

For each batch of *N* question-context pairs, let $(\boldsymbol{q_i}, \boldsymbol{c_i})$ represent positive pairs, where $\boldsymbol{q_i}$ is a question, and $\boldsymbol{c_i}$ is the relevant context. In contrast, $(\boldsymbol{q_i}, \boldsymbol{c_j})$ for i $\neq$ j are treated as negative pairs, where the question $\boldsymbol{q_i}$ is matched with an unrelated context $\boldsymbol{c_j}$. The MNR Loss is designed to minimize the average negative log probability of the positive pairs while treating all other pairs in the batch as negatives.

The objective of the MNR Loss can be formulated as:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{e^{\cos(q_i, c_i)}}{\sum_{j=1}^{N} e^{\cos(q_i, c_j)}}\right) \quad (1)$$

Here, $\boldsymbol{\cos}(\boldsymbol{q_i}, \boldsymbol{c_i})$ denotes the cosine similarity between question $\boldsymbol{q_i}$ and context $\boldsymbol{c_i}$. This loss function forces the model to increase the similarity between positive pairs $(\boldsymbol{q_i}, \boldsymbol{c_i})$ while decreasing it for all other negative pairs within the batch.

Through this fine-tuning process, the embedding model is optimized to more accurately match user queries to relevant contexts, thereby enhancing retrieval performance in RAG Chatbot.

5. EXPERIMENTAL SETTINGS AND EVALUATION

5.1. Settings

All our experiments were implemented on a server with 1x A100 GPU featuring 80GB VRAM. For fine-tuning process, we utilized the *AdamW* optimizer with a *learning rate* of 2e-5 and a *weight decay* of 0.01. The *learning schedule* incorporated a warm-up phase for 10% of the training data using a WarmupLinear scheduler. We set the *batch size* to 16 and performed fine-tuning across 5 *epochs* for all experiments.

In the RAG chatbot setting, we configured the retrieval process to return the top 3 most relevant contexts for each user question. This allows the chatbot to make more informed and accurate responses by considering multiple contextual cues.

5.2. Evaluation scenario and metrics

In this paper, we evaluate our proposed fine-tuning framework through two primary scenarios:

- ***Fine-tuning popular embedding models:*** We compare the retrieval performance of fine-tuned models against their non-fine-tuned models. To assess performance, we use Mean Average Precision at K (MAP@K) with K=100, a commonly applied metric in information retrieval. This metric allows us to thoroughly evaluate the models' capability to retrieve a significant number of relevant documents, which is essential in large-scale datasets where queries often correspond to multiple relevant items.

- ***Evaluating RAG Chatbot Performance***: We deploy a simple RAG Chatbot utilizing the best fine-tuned embedding model in conjunction with Vistral-7B [9], an open-source LLM designed for Vietnamese. We compare this configuration's performance with a RAG Chatbot using OpenAI's embedding model paired with ChatGPT-3.5. The chatbot's performance is evaluated through human verification of the answers, summarized as the correct answer rate (CR), and wrong answer rate (WR) and the decline-to-answer rate (DR).

This dual evaluation strategy allows us to thoroughly assess both the retrieval capabilities of the fine-tuned embedding models and the practical performance of our RAG Chatbot in real-world applications.

6. RESULTS AND DISCUSTION

6.1. Retrieval performance

Table 2 presents the retrieval performance across various embedding models, both before and after fine-tuning, evaluated using the Mean Average Precision at K (MAP@K) metric with K = 100 on the test dataset. The OpenAI text-embedding-small model serves as a baseline for comparison. The improvements observed in our fine-tuned models reveal a substantial enhancement in retrieval capability, with an average MAP@K improvement of 0.1815 across all models.

Notably, our fine-tuning framework also delivered strong improvements for models that were already performing well prior to fine-tuning. The bkai-vietnamese-bi-encoder model, for example, improved from 0.5984 to 0.6325, placing it closer to the OpenAI baseline. Even more impressively, the bge-m3 model, which achieved the highest MAP@K score of 0.7721 after fine-tuning, clearly outperformed the OpenAI model, demonstrating the superior retrieval capabilities of our framework.

English-based models also benefited greatly from the fine-tuning approach. The bge-small-en-v1.5 model saw an improvement from 0.4927 to 0.6115 after fine-tuning, and the colbertv2.0 model, which had a relatively low MAP@K score of 0.1947, increased to 0.5621. These results highlight the generalizability of our fine-tuning approach, which can significantly boost the performance of both Vietnamese and English models.

Table 2. *Retrieval performance (MAP@K) comparison.*

Model	**W/o FT**	**Fine-tuned**
vietnamese-sbert	0.4991	**0.6086**
sup-SimCSE-VietNamese-phobert-base	0.4312	**0.6120**
bkai-vietnamese-bi-encoder	0.5984	**0.6325**
VietnamLegalText-SBERT	0.2753	**0.5174**

bge-small-en-v1.5	0.4927	**0.6115**
colbertv2.0	0.1947	**0.5621**
fpt-vibert-base-cased	0.1880	**0.6049**
vinai-phobert-large	0.3507	**0.5958**
vinai-phobert-base	0.3830	**0.5654**
bge-large-en-v1.5	0.5113	**0.5845**
bge-m3	0.7467	**0.7721**
Average	0.4246	**0.6061** (+0.1815)
OpenAI text-embedding-small	*0.6940*	

By leveraging synthetic data generated by ChatGPT, the framework can quickly generate large volumes of domain-specific data without the need for extensive manual annotation. This is a significant advantage, especially in domain-specific fields like Vietnamese military science, where labeled data is scarce. However, while synthetic data is valuable for generating a large number of question-context pairs, its effectiveness is highly dependent on the quality of questions generated. Although ChatGPT performs well, some questions may lack nuance, potentially limiting the retrieval quality. Additional techniques for filtering or refining synthetic questions could further enhance the framework's performance.

Overall, a key advantage of our fine-tuning framework lies in its ability to significantly improve retrieval performance across diverse embedding models, including both Vietnamese and English models. The average improvement across all models was 0.1815 MAP@K, further highlighting the broad impact of our fine-tuning approach in elevating the retrieval capabilities of both open-source Vietnamese and English embedding models. The performance gains across various embedding models imply that our framework can generalize beyond a single language or domain, potentially enabling cross-domain applications.

6.2. Performance of RAG Chatbot

The performance metrics of various RAG chatbot configurations, focusing on correct answer rate (CR), decline-to-answer rate (DR), wrong answer rate (WR), and GPU VRAM usage for the embedding model, are summarized in table 3.

Table 3. *Performance comparison of different RAG Chatbot configurations.*

RAG Chatbot (Embedding model + LLM)	**Type**	**CR**	**DR**	**WR**	**VRAM GPU (GB)**
OpenAI text-embedding-small + ChatGPT-3.5	Closed source	52.2	19.4	23.8	**0**
Vistral7B + FT bkai-vietnamese-bi-encoder (our)	Open source	68.6	7.5	23.9	1.4
Vistral7B + FT bge-m3 (our)	Open source	**71.6**	**7.4**	**20.9**	7.1

These results demonstrate that our proposed fine-tuning framework significantly enhances the performance of open-source RAG chatbot configurations, clearly outperforming the closed-source OpenAI setup across all key metrics.

In particular, the Vistral7B + FT bkai-vietnamese-bi-encoder configuration, utilizing our fine-tuning framework, achieved a CR of 68.6% and a DR of 7.5%. This represents a substantial improvement over the OpenAI text-embedding-small + ChatGPT-3.5 configuration, which had a CR of 52.2% and a DR of 19.4%. Moreover, the Vistral7B + FT bkai-vietnamese-bi-encoder configuration demonstrated efficient GPU VRAM usage of 1.4 GB, balancing high performance with modest resource consumption. The Vistral7B + FT bge-m3 configuration further achieved the highest CR of 71.6% and the best DR of 7.4%. Although this configuration required 7.1 GB of GPU VRAM, the performance gains in CR and reductions in DR illustrate the effectiveness of fine-tuning with this model. In comparison, the OpenAI configuration, while having no GPU VRAM usage for the embedding model, falls short in terms of CR and has a higher DR. These findings highlight that the closed-source model, despite its lower resource requirements, does not match the performance of our fine-tuned open-source models.

However, it is important to note that the framework's highest-performing model configurations require significant GPU VRAM, which may limit scalability in resource-constrained environments. This resource-intensive nature can be a drawback, particularly for applications aiming to balance high retrieval performance with efficiency.

The reduced DR further suggests that our framework not only enhances retrieval accuracy but also reduces the chatbot's uncertainty, enabling it to respond confidently to a wider range of user queries. This aligns well with the framework's design objective to handle complex and domain-specific data, particularly in fields like military science, where accuracy is crucial.

In summary, the results validate the efficacy of our fine-tuning framework in leveraging open-source models to achieve high-quality retrieval results. The significant improvements in correct answer rate and decline-to-answer rate demonstrate that open-source models, when fine-tuned effectively, can not only match but often exceed the performance of closed-source alternatives, despite potentially higher resource usage.

7. CONCLUSIONS

This study presents a novel approach for enhancing Retrieval-Augmented Generation (RAG) systems, particularly in specialized Vietnamese domains such as military science. By introducing a framework to fine-tune embedding models on synthetic datasets generated by ChatGPT, we have significantly improved the retrieval performance of various models.

Our experimental results reveal that the fine-tuning framework effectively elevates the performance of RAG Chatbot by enhancing the retrieval performance of embedding models. Notably, RAG Chatbot with the fine-tuned models demonstrate substantial improvements in correct answer rate and decline-to-answer rate compared to their non-fine-tuned models and the closed-source OpenAI text embedding model.

In conclusion, this work demonstrates the efficacy of the proposed fine-tuning framework in advancing the state-of-the-art in RAG Chatbot for specialized and low-resource languages. The improvements achieved suggest that similar approaches could be beneficial for other specialized domains and languages, paving the way for more accurate and efficient information retrieval systems in various fields.

REFERENCES

[1]. H. Naveed et al., *"A Comprehensive Overview of Large Language Models,"* (2024), arXiv: arXiv:2307.06435. Accessed: Sep. 09, 2024. [Online]. Available: http://arxiv.org/abs/2307.06435

[2]. N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, *"Large Language Models Struggle to Learn Long-Tail Knowledge,"*, (2023), arXiv: arXiv:2211.08411. Accessed: Sep. 09, 2024. [Online]. Available: http://arxiv.org/abs/2211.08411

[3]. Y. Gao et al., *"Retrieval-Augmented Generation for Large Language Models: A Survey,"* (2024), arXiv: arXiv:2312.10997. Accessed: Sep. 09, 2024. [Online]. Available: http://arxiv.org/abs/2312.10997

[4]. D. Q. Nguyen and A. T. Nguyen, *"PhoBERT: Pre-trained language models for Vietnamese,"*, (2020), arXiv: arXiv:2003.00744. doi: 10.48550/arXiv.2003.00744.

[5]. N. Q. Duc, L. H. Son, N. D. Nhan, N. D. N. Minh, L. T. Huong, and D. V. Sang, *"Towards Comprehensive Vietnamese Retrieval-Augmented Generation and Large Language Models," (*2024), arXiv: arXiv:2403.01616. Accessed: Sep. 09, 2024. [Online]. Available: http://arxiv.org/abs/2403.01616

[6]. *"hmthanh/VietnamLegalText-SBERT · Hugging Face."* Accessed: Sep. 09, 2024. [Online]. Available: https://huggingface.co/hmthanh/VietnamLegalText-SBERT

[7]. M. Tran-Tien, H.-L. Le, D. N. Minh, T. T. Khang, H.-T. Vu, and N. Minh-Tien, *"ViPubmedDeBERTa: A Pre-trained Model for Vietnamese Biomedical Text,"* in Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation, C.-R. Huang, Y. Harada, J.-B. Kim, S. Chen, Y.-Y. Hsu, E. Chersoni, P. A, W. H. Zeng, B. Peng, Y. Li, and J. Li, Eds., Hong Kong, China: Association for Computational Linguistics, (2023), pp. 831–840. Accessed: Sep. 08, 2024. [Online]. Available: https://aclanthology.org/2023.paclic-1.83

[8]. *"OpenAI Platform."* Accessed: Sep. 08, 2024. [Online]. Available: https://platform.openai.com

[9]. *"Viet-Mistral/Vistral-7B-Chat · Hugging Face."* Accessed: Sep. 09, 2024. [Online]. Available: https://huggingface.co/Viet-Mistral/Vistral-7B-Chat

[10]. T. Mikolov, K. Chen, G. Corrado, and J. Dean, *"Efficient Estimation of Word Representations in Vector Space,"* (2013), arXiv: arXiv:1301.3781. doi: 10.48550/arXiv.1301.3781.

[11]. J. Pennington, R. Socher, and C. Manning, *"Glove: Global Vectors for Word Representation,"* in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar: Association for Computational Linguistics, (2014), pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[12]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,"* (2019), arXiv: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[13]. S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie, *"C-Pack: Packaged Resources To Advance General Chinese Embedding,"* (2024), arXiv: arXiv:2309.07597. doi: 10.48550/arXiv.2309.07597.

[14]. O. Khattab and M. Zaharia, *"ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT,"* (2020), arXiv: arXiv:2004.12832. doi: 10.48550/arXiv.2004.12832.

[15]. L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, *"Multilingual E5 Text Embeddings: A Technical Report,"* (2024), arXiv: arXiv:2402.05672. doi: 10.48550/arXiv.2402.05672.

[16]. *"LlamaIndex, Data Framework for LLM Applications."* Accessed: Sep. 09, 2024. [Online]. Available: https://www.llamaindex.ai/

[17]. *"Introduction | Ragas."* Accessed: Sep. 09, 2024. [Online]. Available: https://docs.ragas.io/en/stable/

[18]. M. Henderson et al., *"Efficient Natural Language Response Suggestion for Smart Reply,"* May 01, 2017, arXiv: arXiv:1705.00652. doi: 10.48550/arXiv.1705.00652.

TÓM TẮT

Nâng cao hiệu năng truy xuất của mô hình Embedding thông qua huấn luyện tinh chỉnh trên dữ liệu tạo sinh trong RAG Chatbot cho lĩnh vực Khoa học quân sự Việt Nam

Retrieval-Augmented Generation (RAG) là một công nghệ kết hợp giữa truy xuất thông tin và mô hình ngôn ngữ lớn, cho phép chatbot cung cấp câu trả lời chính xác bằng cách truy vấn các tài liệu liên quan từ kho dữ liệu trước khi tạo ra các phản hồi. Mặc dù RAG chatbot đã cho thấy hiệu quả trong nhiều ứng dụng, nhưng vẫn tồn tại hạn chế trong các lĩnh vực dữ liệu tiếng Việt chuyên ngành, đặc biệt là trong lĩnh vực khoa học quân sự. Để giải quyết thách thức này, bài báo đề xuất một framework để fine-tune các mô hình embedding trên tập dữ liệu tạo sinh bởi ChatGPT nhằm nâng cao hiệu năng truy xuất thông tin trong ứng dụng hỏi đáp lịch sử Viện Công nghệ thông tin (IoIT). Kết quả đánh giá hiệu quả của phương pháp đề xuất trên 11 mô hình embedding phổ biến cho thấy phương pháp đề xuất của chúng tôi cải thiện đáng kể khả năng truy xuất, với trung bình tăng 18,15% chỉ số MAP@K. Chatbot hỏi đáp về lịch sử IoIT, được xây dựng với các mô hình embedding đã fine-tune kết hợp với mô hình ngôn ngữ lớn tiếng Việt Vistral-7B, vượt trội hơn so với các chatbot sử dụng mô hình embedding của OpenAI và ChatGPT. Điều này chứng tỏ tiềm năng cao của công nghệ RAG Chatbot trong việc phát triển các ứng dụng truy xuất thông tin theo ngữ nghĩa trong các lĩnh vực chuyên ngành, đặc biệt là trong lĩnh vực khoa học quân sự.

Từ khoá: Retrieval-augmented generation; Fine-tuning; Dữ liệu tạo sinh; Mô hình ngôn ngữ lớn; Chatbot.