Improving Vietnamese Legal Document Retrieval using Synthetic Data

Pham Tien Son[1], Nguyen Doan Hieu[1], Nguyen Dai An[1], and Dinh Viet Sang[1*]

BKAI Research Center, School of Information and Communication Technology
Hanoi University of Science and Technology, Vietnam
{son.pt204891@sis, hieu.nd231135m, an.nd215296}@sis.hust.edu.vn,
sangdv@soict.hust.edu.vn

Abstract. In the field of legal information retrieval, effective embedding-based models are essential for accurate question-answering systems. However, the scarcity of large annotated datasets poses a significant challenge, particularly for Vietnamese legal texts. To address this issue, we propose a novel approach that leverages large language models to generate high-quality, diverse synthetic queries for Vietnamese legal passages. This synthetic data is then used to pre-train retrieval models, specifically bi-encoder and ColBERT, which are further fine-tuned using contrastive loss with mined hard negatives. Our experiments demonstrate that these enhancements lead to strong improvement in retrieval accuracy, validating the effectiveness of synthetic data and pre-training techniques in overcoming the limitations posed by the lack of large labeled datasets in the Vietnamese legal domain.

Keywords: Information Retrieval · Large Language Models · Natural Language Processing.

1 Introduction

The task of passage retrieval, which involves identifying relevant passages from a large corpus in response to a query, has become increasingly important with the rise of pre-trained language models like BERT [5]. Enhancements such as Sentence-BERT [22] and SimCSE [8] have further improved text embedding techniques, leading to significant advances in Natural Language Processing (NLP).

For Vietnamese legal information retrieval, accurate and efficient retrieval systems are essential, particularly in legal question answering (QA) systems. Prior research has explored various models for this task: [12] introduced a new attention-based architecture, [25] proposed combining BM25 with RoBERTa, and [18] used models like Sentence-BERT and coCondenser to enhance retrieval performance while [19] utilized SimCSE and an ensemble reranker to produce state-of-the-art results on their benchmark data. Despite these efforts, a key limitation remains the scarcity of high-quality annotated datasets in the Vietnamese legal domain, hindering the development of robust retrieval systems.

---

[*] Corresponding author

To address this issue, we propose leveraging large language models (LLMs) to generate synthetic legal queries, creating a substantial dataset to improve passage retrieval. This synthetic data will be used to pre-train and fine-tune models like bi-encoder and ColBERT, enhancing their accuracy and scalability in the Vietnamese legal domain. Our approach not only addresses the immediate issue of data scarcity but also provides a scalable and efficient solution for improving legal information retrieval systems in the Vietnamese legal domain.

The primary contributions of this research are summarized as follows:

1. We present a method for generating synthetic queries based on passages of Vietnamese legal text, resulting in a dataset of 500,000 legal queries and corresponding passages;
2. We implement the 'Query-as-context Pre-training for Dense Passage Retrieval' [26] technique for PhoBERT, further enhancing the retrieval performance of the backbone language model;
3. We demonstrate improvements in retrieval accuracy through the application of bi-encoder and ColBERT retrieval models trained on the newly generated dataset.

The rest of the paper is organized as follows: Section 2 briefly summarizes prior works related to our work. Section 3 describes our method. In Section 4, we will present experimental results on benchmark datasets and compare them with the baseline and state-of-the-art methods. Section 5 analyzes and discusses the results. Finally, we conclude the paper and discuss future work in Section 6.

2   Related Work

Text Embeddings. Recently, there has been a shift of interest towards the use of neural retrieval techniques, which rely on dense vector representations to capture the underlying semantics of both queries and documents. Unlike traditional keyword-based approaches like BM25 or TF-IDF, neural models such as Sentence-BERT [22] and SimCSE [8] produce embeddings that enable semantic matching, allowing for more accurate retrieval based on context rather than exact term matching. These models leverage techniques like [CLS] token pooling or mean-pooling to aggregate embeddings, enabling fast and scalable comparisons via cosine similarity or dot product calculations. The contrastive loss is commonly used during training to fine-tune the models, aligning relevant query-document pairs while distancing irrelevant ones [14].

Multi-vector models like ColBERT [11], extend this approach by representing queries and documents with multiple vectors, providing more detailed interactions for improved retrieval accuracy. One notable development is the M3-Embedding model [4], part of the BGE-M3 project, which supports dense, multi-vector, and sparse retrieval. This model's versatility allows it to adapt to various retrieval scenarios, making it particularly beneficial for applications like legal text analysis. Additionally, M3-Embedding's multilingual capabilities, supporting over 100 languages, and its ability to handle input texts up to 8192 tokens, make it ideal for cross-lingual and large-scale document retrieval tasks.

Cross-encoders, on the other hand, enable even deeper query-document interaction by processing them together in a single transformer model, though their computational intensity limits their use to re-ranking stages in large-scale retrieval systems [24]. Together, these methods represent the evolution of text embeddings in information retrieval, offering more precise and context-aware retrieval mechanisms than their lexical predecessors.

Synthetic data. One of the major obstacles to the widespread adoption of neural retrieval models is their requirement for large supervised training sets to surpass traditional term-based techniques, which are constructed from raw corpora [15]. Thakur et al. [24] find that BM25 remains a robust baseline for out-of-domain tasks. This highlights a critical challenge: in-domain performance cannot reliably predict how well an approach will generalize in a zero-shot setup. Many approaches that outperform BM25 on an in-domain evaluation perform poorly on the BEIR datasets.

Previous works have demonstrated that leveraging pretrained language models to generate queries can significantly enhance the performance of retrieval models in the absence of in-domain labeled data [15,13,3]. Motivated by a line of work on knowledge distillation from black-box LLMs through training on synthetic data generated from them, such as Orca [16] and Phi [9,27] use GPT-3.5/4 [1] to generate query-document pairs across multiple tasks and languages. They then train open-source decoder-only LLMs on this synthetic data using standard contrastive loss, achieving state-of-the-art results in information retrieval benchmarks.

However, using large language models for information retrieval tasks as [27] is costly in practice, both in terms of computational resources and inference latency. Thus, this work examines the prospect of distilling the knowledge from these advanced LLMs into smaller language models such as BERT.

Pre-training tailored for information retrieval. Another approach to improve the performance of dense retrieval that has been studied is enhancing the pre-training process specifically for dense retrieval tasks. Techniques like Condenser [6] involve adding a shallow decoder to the encoder, forcing it to reconstruct masked texts, thereby enhancing the encoder's capacity to produce meaningful embeddings. Based on this, coCondenser [7] further improves performance by incorporating contrastive loss, which ensures that embeddings of similar text spans are brought closer together while those of different spans are pushed apart.

Another innovative approach is CoT-MAE [29], which introduces a contextualized masked auto-encoder structure. This model uses both the target text and its context for reconstruction, further enhancing the encoder's understanding of the text within its broader context. Additionally, models like Query-as-Context [26] extend these ideas by generating queries from passages for pre-training, which has shown promising results in improving retrieval performance. These

advancements underscore the importance of specialized pre-training techniques in boosting the effectiveness of dense retrieval systems.

3 Methodology

3.1 Overall

In this subsection, we outline our comprehensive workflow for generating synthetic query datasets and fine-tuning retrieval models using Vietnamese legal texts. This workflow is visually represented in Figure 1, which illustrates the sequential stages from data collection to model fine-tuning.
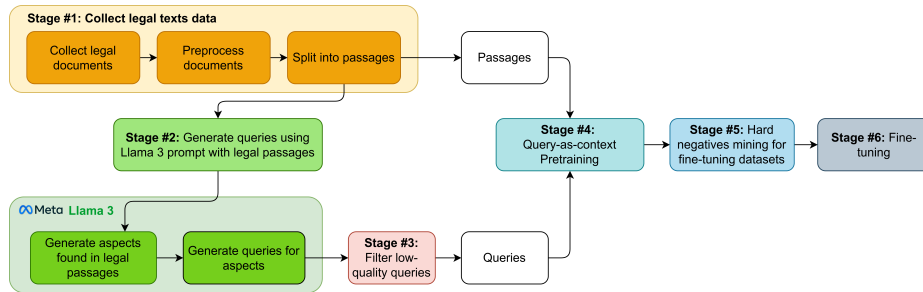

Fig. 1. Workflow for generating synthetic queries and fine-tuning retrieval models using Vietnamese legal texts.

Our methodology is divided into several key stages:

Stage 1: Collect legal text data. This initial stage involves the collection and preprocessing of legal documents. The documents are split into smaller passages suitable for further processing.

Stage 2: Generate queries using Llama 3 prompt with legal passages. Using the Meta Llama 3 model, we generate synthetic queries based on aspects identified in the legal text passages. This involves generating aspects from the legal texts and then crafting queries for these aspects.

Stage 3: Filter low-quality queries. We remove low-quality queries that explicitly refer to the input passage and those that are only shallowly relevant to the input passage. For queries that are only shallowly relevant, we use the BGE-M3 dense retriever to filter out synthetic queries that cannot recover their input passage within the top 40 retrieved results.

Stage 4: Query-as-Context Pre-training. We employ the generated queries to further pre-train our language model, focusing on enhancing its ability to understand and retrieve relevant passages.

Stage 5: Hard negatives mining for fine-tuning datasets. We mine hard negative examples to create a robust fine-tuning dataset, which is crucial for improving the retrieval model's accuracy.

Stage 6: Fine-tuning. The final stage involves fine-tuning the pre-trained model with the generated data, optimizing it for the specific task of legal text retrieval.

3.2 Data Curation

Wang et al. [27] use GPT-3.5/4 to generate queries along with corresponding positive and hard negative passages by maintaining output diversity through two stages of generation: first generating retrieval tasks, then using them to generate query-passage pairs. As our objective focuses on creating a domain-specific dataset for Vietnamese legal text retrieval, relying solely on prompt engineering for this task would be complex and inefficient. Therefore, we opted to use collected legal text passages as input for an LLM to generate queries related to the content of each passage.

Our main source for collecting legal documents was thuvienphapluat.vn. The website hosts a wide range of legal documents, including Laws, Decrees, Circulars, Joint Circulars, Resolutions, Ordinances, Decisions, and the Constitution. After scraping both the metadata and full-text content, we preserved the hierarchical structure of each document, which typically includes chapters, sections, articles, and clauses. This structure was essential for maintaining context and ensuring that the text remained coherent after being split into smaller passages. Each passage retained crucial information, such as the document's domain, title, header, and main content, which provided the necessary context for accurate query generation. This process resulted in 143,261 passages, which were used as input for the LLM to generate high-quality, contextually relevant queries.

3.3 Synthetic Query Generation

For generating synthetic queries, we chose the open-source LLM Llama 3 70B [2] due to its strong performance, particularly in Vietnamese. Llama 3, trained on over 15 trillion tokens, outperforms many LLMs with a similar parameter count and demonstrates robust multilingual capabilities, making it well-suited for our needs.

To maintain diversity and relevance in query generation, we experimented with different prompting techniques. A direct approach, where the model generated questions without identifying distinct aspects of the passage, often led to less diverse and sometimes irrelevant queries. Through various prompt designs, we discovered that instructing the model to identify 1 - 5 different aspects covered in the passage and then generate a question for each aspect yielded the most relevant and diverse queries. In Figure 2, we show the prompt template used to generate synthetic queries from legal text passages. Examples of a generated query and its corresponding passage are shown in Table 1. A quantitative analysis of the generation method will be later discussed in Section 5.1.

Applying this method, we generated over 620,000 legal queries from 140,292 passages extracted from our curated Vietnamese legal text collection. We then employed the BGE-M3 dense retriever [4], which demonstrated strong zero-shot

You are an advanced legal query generator with specialized skills in analyzing legal documents. When provided with an excerpt from a legal document, your task is to identify 1-5 critical aspects or implications that might interest or impact the readers. These aspects should address various dimensions of the content, focusing on rights, obligations, potential legal issues, or general legal awareness, exclusively within provided grounded content. Do not consider information in document's source for this analysis. The following is the mentioned excerpt:

…

For each identified critical aspect, generate a single question. These questions should reflect plausible inquiries that an average citizen might have, relating directly to the document but formulated in a manner accessible to someone unfamiliar with the presence of the legal text or information being asked about. Phrase the questions as if coming from a layperson who has not read or seen the legal text ever.

Your output should be in JSON format, listing the critical aspects identified and a corresponding question for each aspect. Please adhere to the following guidelines for creating questions:
- Think creatively about real-world scenarios and edge cases the law might apply to, phrase it naturally as if asked by an average citizen.
- The queries should be ones that could reasonably be answered by the information exclusively within provided grounded content only. Do not ask information in document's source.
- Each query should be one sentence only and its length is no more than 120 words.
- Try to phrase each of the question as detailed as possible, as if you haven't never seen the legal text and are trying to looking for it using keywords in the question, you may need to include details in document's source and domain for this aim. You should not quote the exact legal text code (like 02/2017/TT-BQP). The better way is to include information on the content of document as in document's source instead like the executive body published the document (e.g. "Bộ Y tế quy định thế nào về ..."). In the case you have to refer to the legal text, use words like: "Quy định pháp luật", "Pháp luật", "Luật". Don't use the word "này".
- Present your analysis and questions in Vietnamese.

…

Structure your output in the JSON format below:
```

{
  "aspects": [
    [Brief description of the aspect 1],
    [Brief description of the aspect 2],
     ...
  ],
  "questions": [
    [Your question related to aspect 1 of the legal text],
    [Your question related to aspect 2 of the legal text],
     ...
  ]
}
```

Ensure to replace the placeholders with actual analysis and questions based on the legal text provided, and in Vietnamese. Answer with the JSON and nothing else.

Response:

Fig. 2. The shortened prompt template we used to generate synthetic queries from legal text passages, with placeholders for input documents and few-shot examples omitted.

performance in our testing, to filter out queries whose corresponding passages did not appear in the top 40 relevant results. Additionally, we excluded queries that directly referred to the passage using terms like "quy định này" or "thông tư này". This process results in a final dataset of 507,152 Vietnamese legal queries, covering a wide range of legal domains, which is then used to pre-train and fine-tune our retrievers.

We generated these synthetic queries using Llama 3 70B through Together AI's free credits program, with the entire generation process costing approximately 200 USD.

Table 1. Example of a generated query-passage pair for the domain "Tiền tệ - Ngân hàng".

Domain	Tiền tệ - Ngân hàng
Header	Mục 1. CHUẨN BỊ THANH TRA, Chương II. TRÌNH TỰ, THỦ TỤC TIẾN HÀNH CUỘC THANH TRA THEO KẾ HOẠCH THANH TRA, Thông tư 36/2016/TT-NHNN quy định về trình tự, thủ tục thanh tra chuyên ngành Ngân hàng do Thống đốc Ngân hàng Nhà nước Việt Nam ban hành.
Content	5. Trưởng đoàn thanh tra tổ chức họp Đoàn thanh tra để phổ biến kế hoạch tiến hành thanh tra được duyệt và phân công nhiệm vụ cho các Tổ thanh tra, Nhóm thanh tra, các thành viên của Đoàn thanh tra; thảo luận, quyết định về phương pháp, cách thức tổ chức tiến hành thanh tra; sự phối hợp giữa các thành viên Đoàn thanh tra, các cơ quan, đơn vị có liên quan trong quá trình triển khai thanh tra. Trong trường hợp cần thiết, người ra quyết định thanh tra hoặc người được người ra quyết định thanh tra ủy quyền dự họp và quán triệt mục đích, yêu cầu, nội dung thanh tra và nhiệm vụ của Đoàn thanh tra. Việc phân công nhiệm vụ cho các Tổ thanh tra, Nhóm thanh tra, các thành viên Đoàn thanh tra phải thể hiện bằng văn bản. 6. Tổ trưởng thanh tra, Nhóm trưởng thanh tra, thành viên Đoàn thanh tra xây dựng kế hoạch thực hiện nhiệm vụ được phân công và báo cáo Trưởng đoàn thanh tra trước khi thực hiện thanh tra tại tổ chức tín dụng.
Aspect 1	Trách nhiệm của Trưởng đoàn thanh tra trong việc tổ chức và phân công nhiệm vụ
Query 1	Ngân hàng Nhà nước quy định Trưởng đoàn thanh tra phải làm gì để chuẩn bị cho cuộc thanh tra?
Aspect 2	Quy trình xây dựng và báo cáo kế hoạch thực hiện nhiệm vụ của các Tổ thanh tra, Nhóm thanh tra
Query 2	Khi được phân công nhiệm vụ, các Tổ thanh tra, Nhóm thanh tra phải làm gì để chuẩn bị cho cuộc thanh tra?

3.4 Pre-training

Query-as-Context pre-training [26] is based on the observation that text spans within the same document can vary significantly in semantics, potentially weakening the effectiveness of traditional pre-training techniques. However, one limitation noted in this approach is that the T5 model used often produced a substantial number of unrelated query-passage pairs, which could diminish its overall effectiveness. This aligns well with the first part of our work — generating a legal query dataset — where we address this issue by utilizing an advanced LLM to produce queries and implementing a filtering process to remove irrelevant pairs, as detailed in the previous subsection.

Using the pairs of collected passages x_i and corresponding legal queries y_i generated by Llama 3, we apply the loss function from the Contextualized Masked Autoencoder (CoT-MAE) framework [29]. Specifically, the encoder reconstructs the passage x_i using its unmasked tokens, while the decoder reconstructs the query y_i by leveraging both its unmasked tokens and the contextual passage x_i. The training objective combines the encoder's masked language modeling (MLM) loss and the decoder's context-supervised MLM loss, ensuring that the model learns to effectively integrate the query and passage context during pre-training.

Table 2. Fine-tuning dataset statistics.

Training Dataset	Translated	#Passage	#Query	#Positive Pairs
Zalo Legal 2021 (80%)		61,425	2,556	2,556
MS-MARCO [17]	✓	8,841,823	502,939	532,751
SQuAD 2.0 [21]	✓	13,317	60,942	60,942
TVPL		224,006	165,334	189,641

3.5   Fine-tuning

To verify the effectiveness of our pre-training, we fine-tuned a bi-encoder and ColBERT model on downstream retrieval tasks. Our fine-tuning process is based on a single-stage pipeline with hard negative mining, utilizing a comprehensive set of datasets.

For both models, we utilized the MS-MARCO passage ranking dataset, SQuAD 2.0, 80% of the Legal Text Retrieval Zalo 2021 training challenge dataset, and the dataset collected from thuvienphapluat.vn. Additionally, our synthetic query data generated for pre-training is also used in this fine-tuning process. For all datasets, we create hard negative passages for each training query from the BGE-M3 dense retrieval model.

For each query q^+, the positive passage p^+ forms a pair (h_{p^+}, h_{q^+}). The negative samples $\{p^-\}$ included hard negatives identified by the BGE-M3 dense retrieval and in-batch negative passages. The training objective for both models is to maximize the similarity between the query and the positive passage while minimizing the similarity between the query and the negative passages. This is achieved using the InfoNCE loss function:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(h_{q^+}, h_{p^+})/\tau)}{\sum_{p \in \{p^+, p^-\}} \exp(\text{sim}(h_q, h_p)/\tau)},$$

where τ is a temperature hyper-parameter fixed to 1, and $\text{sim}(\cdot, \cdot)$ represents the dot product similarity function.

4   Experiment

4.1   Datasets

Table 2 presents the datasets used for fine-tuning and evaluation. Our primary fine-tuning datasets include MS-MARCO [17], SQuAD 2.0 [21], 80% of the training set from the Legal Text Retrieval Zalo 2021 challenge (Legal Zalo 21), and the newly introduced TVPL dataset. Since MS-MARCO and SQuAD 2.0 are originally in English, we translated them into Vietnamese using Google Translate, following the approach of [20]. The use of large, translated datasets has proven beneficial for improving the performance of monolingual retriever models due to the absence of comparably large annotated datasets for Vietnamese.

We developed TVPL — a new benchmark dataset for Vietnamese legal text retrieval. TVPL is named after the thuvienphapluat.vn website, from which its legal QA articles were sourced. The dataset comprises a training set of 165,334 queries and a test set of 10,000 queries, along with a corpus of 224,006 legal passages. This dataset addresses previous limitations of scale and diversity, providing a more comprehensive benchmark for evaluating retrieval models in the Vietnamese legal domain.

Additionally, our 507,152 generated queries using the Llama3-70B model based on passages extracted from Vietnamese legal texts are also used in this fine-tuning stage. This synthetic dataset spans a broad range of legal domains, as shown in Figure 3.
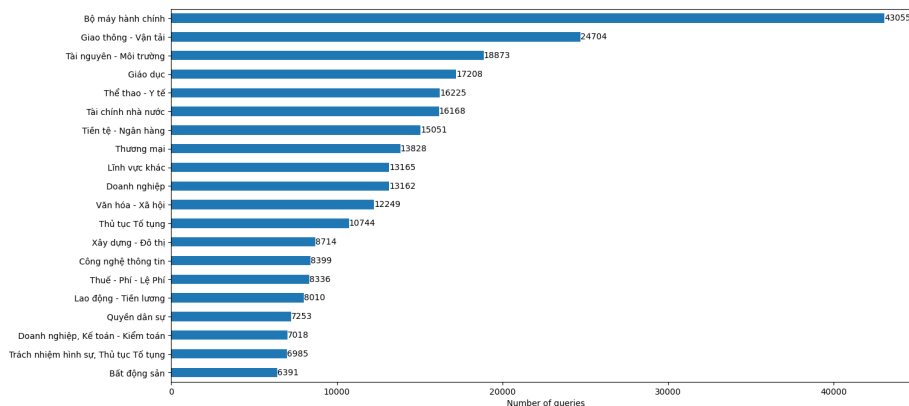

Fig. 3. Top 20 domains by number of queries in synthetic query dataset.

For evaluation, we used 20% of the Legal Zalo 21 dataset and 10,000 test queries from TVPL for in-domain testing. Additionally, we employed the Vietnamese Wiki Question Answering dataset from the Zalo AI Challenge 2019 for out-of-domain evaluation.

4.2   Model Pre-training and Fine-tuning

We used the synthetic dataset for pre-training, with text segmented by the Underthesea library. During pre-training, we select a batch of passages at each step and randomly choose a candidate query as context for each passage to form a relevant pair. The encoder for CoT-MAE was initialized with a pretrained PhoBERT-base-v2 model, while the decoder was trained from scratch. Pre-training was conducted for 2,000 steps using the AdamW optimizer with a learning rate of 1e-4, a batch size of 1024, and a linear schedule on a TPU v4-8. After training, the decoder was discarded, and only the encoder was retained for fine-tuning.

Table 3. Performance comparison on our TVPL and Legal Zalo 21 benchmarks.

Model	TVPL				Legal Zalo 21			
	MRR@10	MAP@10	R@10	R@100	MRR@10	MAP@10	R@10	R@100
Sparse retrieval								
BM25	21.60	20.87	41.11	70.64	51.53	31.68	48.43	72.33
Dense retrieval								
vietnamese-sbert [10]	48.93	45.89	74.37	92.87	46.31	32.74	48.95	71.88
vietnamese-bi-encoder [20]	48.38	46.42	68.92	86.44	80.69	56.65	66.93	80.11
mE5$_{base}$ [28]	19.39	18.43	33.50	58.69	59.72	54.25	**71.77**	**84.05**
BGE-M3 [4]	32.68	31.32	51.78	74.46	64.43	44.02	57.28	75.28
Bi-encoder	70.37	67.96	87.09	96.34	79.31	57.90	68.02	81.76
CoT-MAE Bi-encoder	**70.69**	**68.25**	**87.34**	**96.92**	**80.03**	**58.41**	**69.08**	81.61
ColBERT	73.90	71.39	88.68	96.46	84.15	60.54	67.93	80.84
CoT-MAE ColBERT	**74.61**	**72.04**	**89.29**	96.41	84.08	**60.76**	**69.17**	**81.34**

For fine-tuning, we utilized the four annotated datasets presented in the previous subsections along with our generated synthetic data. Legal Zalo 21 passages were chunked to PhoBERT's 256-token limit. Hard negatives for each query were mined using the BGE-M3 dense retrieval model. The bi-encoder model was trained with a batch size of 64 across 4 TPU chips, optimized with AdamW at a 2e-5 learning rate for 170,000 steps (5 epochs), pairing each query with one positive passage and 7 hard negatives. ColBERT was trained with a batch size of 16, 15 hard negatives, and a longer training period of 290,000 steps (9 epochs), as the model continued improving throughout. ColBERT embeddings were compressed to 2 bits per dimension for evaluations.

4.3 Baselines

For evaluation, our baseline methods include both sparse and dense retrieval approaches, as outlined in Table 3. The sparse retrieval baseline is represented by BM25. For dense retrieval baselines, we include results from the monolingual models vietnamese-sbert [10] and vietnamese-bi-encoder [20], as well as the multilingual models BGE-M3 [4] and mE5$_{base}$ [28].

4.4 Main Results

As shown in Table 3, the results demonstrate that fine-tuning with additional generated data improves retrieval performance across all evaluation metrics. Pre-training further enhances these results, with notable improvements in both the bi-encoder and ColBERT models. ColBERT, with its more granular multi-vector retrieval, achieves the highest scores across both the TVPL and Legal Zalo 21 benchmarks.

Table 4. Out-of-domain evaluation on Vietnamese Wiki Question Answering dataset from the Zalo AI Challenge 2019.

Model	Zalo QA 19			
	MRR@10	MAP@10	R@10	R@100
Sparse retrieval				
BM25	45.33	42.54	67.56	87.17
Dense retrieval				
vietnamese-sbert [10]	48.93	45.89	74.37	92.87
vietnamese-bi-encoder [20]	68.15	64.81	85.39	96.20
mE5$_{base}$ [28]	**72,76**	**70.03**	**91.55**	**98.57**
BGE-M3 [4]	**76.69**	**74.37**	**94.26**	**99.06**
Bi-encoder	69.57	66.36	86.60	96.91
CoT-MAE Bi-encoder	68.22	64.91	85.38	96.20
ColBERT	70.34	67.58	88.72	96.86
CoT-MAE ColBERT	**72.38**	**69.72**	**90.47**	**97.71**

4.5   Out-of-domain Evaluation

To further evaluate the robustness and generalization capabilities of our models, we conducted an out-of-domain evaluation on the Vietnamese Wiki Question Answering dataset from the Zalo AI Challenge 2019. This dataset contains query-passage pairs covering a wide range of topics beyond the legal domain, on which our models were specifically pre-trained and fine-tuned.

Despite being trained solely in the legal context, our models demonstrated improved performance on the out-of-domain dataset, as shown in Table 4. Notably, the pre-trained and fine-tuned ColBERT model achieves scores close to those of larger multilingual retrievers, such as mE5$_{base}$ and BGE-M3. These models, with significantly larger parameter counts (mE5$_{base}$ at 270M and BGE-M3 at 560M, compared to our models with 124M), are still strong candidates for zero-shot retrieval tasks. The results suggest that our approach, though specialized for legal text, maintains strong generalization capabilities in broader retrieval scenarios.

5   Analyses

5.1   Effects of Aspect-guided Query Generation Prompt

We analyze the improvements brought by the aspect-guided query generation method (Section 3.3) compared to basic prompting, where the LLM generates queries directly from input passages. Performance is evaluated using passage hit rate (the percentage of queries retrieving their corresponding passage) and document hit rate (the percentage of queries retrieving the correct document). We use the BGE-M3 dense retriever [4] to rank the top-k relevant passages for each query.

Table 5. Passage hit rate and document hit rate for different top-k values of 10.000 queries generated by Llama 3 using two prompting methods.

Prompt	$k = 10$		$k = 20$		$k = 40$	
	Passage hit	Document hit	Passage hit	Document hit	Passage hit	Document hit
Basic prompt	8.26	87.23	11.93	91.05	16.15	93.38
Aspect-guided query generation prompt	**82.06**	**91.60**	**87.92**	**94.51**	**91.90**	**96.30**

Table 6. Impact of compression on storage and retrieval performance on TVPL benchmark.

Index	Storage (MB)	F2@10	MRR@10	MAP@10	Recall@10	Recall@100
ColBERT						
1 bit	647.33	0.3371	73.61	71.04	88.49	96.24
2 bits	1094.04	0.3406	74.61	72.04	89.29	96.41
4 bits	1987.48	0.3411	74.93	72.34	89.43	96.51
8 bits	3774.34	0.3407	75.02	72.43	89.31	96.49
Bi-encoder	672.02	0.3407	70.69	68.25	87.34	96.92

As shown in Table 5, the aspect-guided method significantly outperforms basic prompting across different top-k values (10, 20, 40). At $k = 10$, it achieves an 82.06% passage hit rate and 91.60% document hit rate, compared to 8.26% and 87.23%, respectively, for basic prompting.

5.2 Quality - Space Footprint Trade-off: ColBERT's Residual Compression

We examine the impact of increasing the number of bits for compression on both retrieval accuracy and storage requirements. Experiments were conducted on 224,006 passages from the TVPL dataset, with evaluation metrics on its test set. ColBERT's residual compression, as proposed in [23], offers improvements in both storage and retrieval performance. While higher bit sizes improve performance slightly, they come with a significant increase in storage. In contrast, The 1-bit ColBERT configuration performs competitively with minimal storage (647.33MB), even outperforming bi-encoder (672.02MB) in metrics like MRR@10 and MAP@10.

6   Conclusion and Future Work

In this work, we proposed methods to improve Vietnamese legal text retrieval using synthetic data. Our key contributions include generating synthetic legal queries from Vietnamese legal text passages with a pre-trained LLM, creating a dataset of 500K query-passage pairs, and significantly enhancing retrieval accuracy with bi-encoder and ColBERT models trained on this dataset. Our

experiments demonstrate the effectiveness of fine-tuning with synthetic data for improving model performance. We also applied the query-as-context CoT-MAE pre-training technique, which further boosted retrieval accuracy. The combination of synthetic data and CoT-MAE pre-training consistently yielded superior performance in both in-domain and out-of-domain evaluations.

The dataset generated in this work has been made publicly available on Hugging Face under the CC BY 4.0 license[1]. Additionally, we are also publishing the TVPL benchmark dataset queries[2]. We hope that these datasets and the method introduced can help advance the development of new language models for Vietnamese, potentially extending beyond legal contexts, and support applications like legal search and question-answering systems to benefit both public servants and citizens.

In future work, we want to explore the potential of using generated queries as inputs for large language models to create an artificial legal corpus. This corpus could, in turn, produce additional queries, creating a cycle to expand the dataset further. A deeper qualitative analysis comparing synthetic data to real-world data is necessary to understand this approach's strengths and limitations better. We also aim to examine potential performance collapse in models primarily trained on synthetic data and investigate strategies to mitigate this risk.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023)
2. AI@Meta: Llama 3 Model Card (2024), `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`
3. Bonifacio, Luiz and Abonizio, Hugo and Fadaee, Marzieh and Nogueira, Rodrigo: Inpars: Unsupervised dataset generation for information retrieval. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2387–2392. Association for Computing Machinery (2022)
4. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 2318–2335. Association for Computational Linguistics (Aug 2024)

---

[1] `https://huggingface.co/datasets/phamson02/large-vi-legal-queries`
[2] `https://huggingface.co/datasets/phamson02/tuvanphapluat`

5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019)

6. Gao, L., Callan, J.: Condenser: a Pre-training Architecture for Dense Retrieval. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021. pp. 981–993. Association for Computational Linguistics (2021)

7. Gao, L., Callan, J.: Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022. pp. 2843–2853. Association for Computational Linguistics (2022)

8. Gao, T., Yao, X., Chen, D.: SimCSE: Simple Contrastive Learning of Sentence Embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6894–6910. Association for Computational Linguistics (Nov 2021)

9. Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C.C.T., Giorno, A.D., Gopi, S., Javaheripi, M., Kauffmann, P.C., de Rosa, G.H., Saarikivi, O., Salim, A., Shah, S., Behl, H., Wang, X., Bubeck, S., Eldan, R., Kalai, A.T., Lee, Y.T., Li, Y.: Textbooks are all you need (2024), https://openreview.net/forum?id=Fq8tKtjACC

10. keepitreal: Vienamese sbert. https://huggingface.co/keepitreal/vietnamese-sbert, online; accessed 18 September 2024

11. Khattab, O., Zaharia, M.: 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. p. 39–48. Association for Computing Machinery (2020)

12. Kien, P.M., Nguyen, H.T., Bach, N.X., Tran, V., Le Nguyen, M., Phuong, T.M.: Answering legal questions by learning neural attentive text representation. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 988–998 (2020)

13. Liang, D., Xu, P., Shakeri, S., dos Santos, C.N., Nallapati, R., Huang, Z., Xiang, B.: Embedding-based Zero-shot Retrieval through Query Generation. CoRR **abs/2009.10270** (2020)

14. Lin, S.C., Asai, A., Li, M., Oguz, B., Lin, J., Mehdad, Y., Yih, W.t., Chen, X.: "How to Train Your Dragon: Diverse Augmentation Towards Generalizable Dense Retrieval". In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 6385–6400. Association for Computational Linguistics (Dec 2023)

15. Ma, J., Korotkov, I., Yang, Y., Hall, K., McDonald, R.: "Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation". In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 1075–1088. Association for Computational Linguistics (Apr 2021)

16. Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., Awadallah, A.: Orca: Progressive Learning from Complex Explanation Traces of GPT-4. CoRR **abs/2306.02707** (2023)

17. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human generated machine reading comprehension dataset. In: Besold, T.R., Bordes, A., d'Avila Garcez, A.S., Wayne, G. (eds.) Proceedings of the

Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016). vol. 1773. CEUR-WS.org (2016)

18. Pham, N., Nguyen, H., Do, T.: Multi-stage Information Retrieval for Vietnamese Legal Texts. CoRR **abs/2209.14494** (2022)

19. Pham Duy, A., Le Thanh, H.: A Question-Answering System for Vietnamese Public Administrative Services. In: Proceedings of the 12th International Symposium on Information and Communication Technology. pp. 85–92 (2023)

20. Quang Duc, N., Hai Son, L., Nhan, N.D., Dich Nhat Minh, N., Thanh Huong, L., Viet Sang, D.: Towards Comprehensive Vietnamese Retrieval-Augmented Generation and Large Language Models. arXiv e-prints (Mar 2024)

21. Rajpurkar, P., Jia, R., Liang, P.: Know What You Don't Know: Unanswerable Questions for SQuAD. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 784–789. Association for Computational Linguistics (Jul 2018)

22. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 2019. pp. 3980–3990. Association for Computational Linguistics (2019)

23. Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., Zaharia, M.: ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022. pp. 3715–3734. Association for Computational Linguistics (2022)

24. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)

25. Van, H.N., Nguyen, D., Nguyen, P.M., Le Nguyen, M.: Miko team: Deep learning approach for legal question answering in alqac 2022. In: 2022 14th International Conference on Knowledge and Systems Engineering (KSE). pp. 1–5. IEEE (2022)

26. W, X., Ma, G., Qian, W., Lin, Z., Hu, S.: Query-as-context pre-training for dense passage retrieval. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 1906–1916. Association for Computational Linguistics (Dec 2023)

27. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., Wei, F.: Improving text embeddings with large language models. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11897–11916. Association for Computational Linguistics (Aug 2024)

28. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., Wei, F.: Multilingual E5 Text Embeddings: A Technical Report. CoRR **abs/2402.05672** (2024)

29. Wu, Xing and Ma, Guangyuan and Lin, Meng and Lin, Zijia and Wang, Zhongyuan and Hu, Songlin: Contextual masked auto-encoder for dense passage retrieval. In: Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence. AAAI Press (2023)