Improving Vietnamese-English Cross-Lingual Retrieval for Legal and General Domains

Toan Ngoc Nguyen[1]*, Nam Le Hai[1]*, Nguyen Doan Hieu[1]*, Dai An Nguyen[1],
Linh Ngo Van[1], Thien Huu Nguyen[2], Sang Dinh[1]†

[1]BKAI Research Center, Hanoi University of Science and Technology, [2]University of Oregon

Abstract

Document retrieval plays a crucial role in numerous question-answering systems, yet research has concentrated on the general knowledge domain and resource-rich languages like English. In contrast, it remains largely underexplored in low-resource languages and cross-lingual scenarios within specialized domain knowledge such as legal. We present a novel dataset designed for cross-lingual retrieval between Vietnamese and English, which not only covers the general domain but also extends to the legal field. Additionally, we propose auxiliary loss function and symmetrical training strategy that significantly enhance the performance of state-of-the-art models on these retrieval tasks. Our contributions offer a significant resource and methodology aimed at improving cross-lingual retrieval in both legal and general QA settings, facilitating further advancements in document retrieval research across multiple languages and a broader spectrum of specialized domains. All the resources related to our work can be accessed at huggingface.co/datasets/bkai-foundation-models/crosslingual.

1 Introduction

Document retrieval systems play a crucial role in question-answering (QA) frameworks by identifying relevant documents that provide the necessary information to answer a given query. However, the majority of existing document retrieval systems (Karpukhin et al., 2020; Khattab and Zaharia, 2020; Gao et al., 2021; Sachan et al., 2022; Dong et al., 2023) and datasets (Nguyen et al., 2016; Kwiatkowski et al., 2019; Thakur et al., 2021; Qiu et al., 2022; Muennighoff et al., 2023) are designed to operate within a single language, typically targeting resource-rich languages like English or Chinese. This monolingual focus limits the effectiveness of these systems in multilingual contexts, where users may pose queries in one language while the relevant documents are in another.

Some studies have tried to explore cross-lingual information retrieval (Liu et al., 2020; Bonab et al., 2020; Huang et al., 2023; Louis et al., 2024), yet these efforts have largely concentrated on high-resource languages and general domain knowledge, leveraging extensive resources and pre-existing knowledge bases. Meanwhile, Vietnamese remains largely underexplored in this context, primarily due to the limited availability of datasets necessary for pretraining and fine-tuning representation models in this language. For example, Vietnamese accounts for less than 1% of the total pretraining data in the BGE M3 model (Chen et al., 2024). Additionally, general domain datasets (Nguyen et al., 2016; Thakur et al., 2021; Muennighoff et al., 2023) are frequently derived from open-domain sources such as Wikipedia, web documents, or news articles, that typically involve quite short documents. While this is valuable for general QA, it fails to address the complexities of specialized domains, where documents are often lengthy and domain-specific, such as legal documentation. Consequently, there is a need to develop cross-lingual document retrieval systems that can effectively handle low-resource languages and specialized domains, ensuring more comprehensive and context-aware QA solutions.

To address these gaps, we present a novel benchmark aimed at evaluating cross-lingual information retrieval (CLIR) between Vietnamese and English. In addition to general knowledge question answering, our dataset enables the investigation of retrieval systems within the legal domain using the Vietnamese Law Library. From this resource, we develop a retrieval model that demonstrates strong performance across both general knowledge and legal domains.

In summary, our contributions are as follows:

*Equally contributed.

†Corresponding author: sangdv@soict.hust.edu.vn

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 142–153
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

1. **Low-Resource Legal Dataset:** We introduce VNLawQC, a dataset designed to explore information retrieval in the legal domain in Vietnamese, along with VNSynLawQC, a synthetic dataset generated by large language models (LLMs) to further augment the training data.

2. **Cross-Lingual Legal Retrieval Dataset:** To enable cross-lingual retrieval, we construct a Vietnamese-English dataset that supports both general and legal domain knowledge. This dataset is constructed using translation models, followed by careful filtering to ensure the selection of high-quality data.

3. **Novel Methodologies for CLIR:** We propose an Auxiliary loss function and Symmetrical training procedure that demonstrates significant improvement in cross-lingual information retrieval scenarios across general knowledge and legal domains.

2 Related Work

Recently, Information Retrieval has attracted considerable attention, with document retrieval emerging as one of the central focuses. Several methods have been proposed to address this task, which can generally be classified into three approaches: dense retrieval (Karpukhin et al., 2020; Xiong et al., 2021; Wang et al., 2022), lexical retrieval (Dai and Callan, 2020; Gao et al., 2021), and multi-vector retrieval (Khattab and Zaharia, 2020; Chen et al., 2024). Numerous datasets have also been developed to evaluate these systems (Nguyen et al., 2016; Thakur et al., 2021; Muennighoff et al., 2023).

However, these methods and datasets primarily focus on monolingual scenarios. Recently, several studies have explored cross-lingual settings, where queries and documents are in different languages (Liu et al., 2020; Bonab et al., 2020; Huang et al., 2023; Louis et al., 2024). In contrast, some prior studies have investigated specific domains, such as the legal, extending beyond general knowledge QA, but still focusing on monolingual scenarios (Sugathadasa et al., 2019; Louis and Spanakis, 2022; Sansone and Sperlí, 2022; Nguyen et al., 2024; Su et al., 2024).

3 Methodology

3.1 Dataset Construction

Data Construction Pipeline: Figure 1 illustrates the complete pipeline for constructing our dataset for cross-lingual information retrieval (CLIR) between Vietnamese and English. Overall, our efforts concentrate on collecting data from the Vietnamese legal domain, where resources are limited while utilizing existing datasets from both general and legal fields to generate cross-lingual data through translation approaches. Moreover, to prevent data leakage, we implement data deduplication across the legal datasets using the MinHash technique (Luo et al., 2015; Zhu and Markovtsev, 2017).

Dataset	Language	Train	Eval	Corpus
MS-MARCO (Nguyen et al., 2016)	en	457,361	0	8,841,823
SQuADv2 (Rajpurkar et al., 2018)	en	60,942	0	13,317
ZaloLegal2021 (Zalo AI Team, 2021)	vi	2,556	640	61,060
ZaloWikipediaQA (Zalo AI Team, 2019)	vi	0	4,399	15,957
VNLawQC	vi	165,347	9,992	224,008
VNSynLawQC	vi	503,068	0	140,291

Table 1: The original language, number of training and evaluation samples, and the corpus size for each dataset. *en* refers to English, while *vi* denotes Vietnamese. *Corpus* denotes the total number of documents in the dataset.

Legal Retrieval Dataset: We introduce VNLawQC sourced from Vietnamese Law Library[1] (VLL). The VLL contains articles that address questions spanning multiple aspects of the legal domain. Each article provides an answer supported by one or more legal documents, with hyperlinks directing to the corresponding documents. To create the VNLawQC dataset, we constructed query-passage pairs based on the structure of these articles. Specifically:

1. The queries were extracted directly from the questions presented in the articles.

2. For the passages, we followed the hyperlinks in each answer to access the referenced legal documents. The relevant sections from these documents were then extracted to serve as the passages.

After parsing content from HTML tags, we apply basic cleaning techniques, including capitalizing legal terms (e.g. “*Điều*”–“*Article*”, “*Khoản*”–“*Clause*”), normalizing Unicode characters, and standardizing tone marks, following prior works on Vietnamese text processing (Vu et al., 2018; Nguyen and Nguyen, 2020). As a result, the VNLawQC dataset is composed of query-passage pairs, where each query can have multiple associated passages if the answer references multiple

[1] https://thuvienphapluat.vn

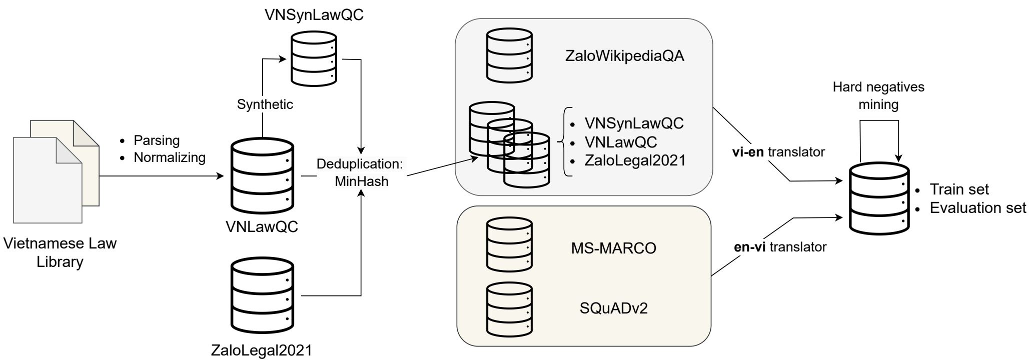

Figure 1: An Overview of the Our Data Construction Pipeline.

legal documents. This process ensures that the dataset captures a realistic mapping of questions to relevant legal information.

Synthetic Legal Retrieval Dataset: We used LLama-3-70B model (Dubey et al., 2024) to generate synthetic question-context pairs from 140K distinct passages in VNLAWQC, creating VNSYNLAWQC to augment the training data (detailed in Appendix B). Llama-3-70B was chosen for its strong performance, especially in languages like Vietnamese, and its ability to generate high-quality queries. A key challenge in using LLMs for query generation is balancing diversity and relevance. We experimented with different prompt techniques and found that instructing the model to identify 1-5 aspects in a passage and generate a question for each aspect resulted in the most relevant and diverse queries (see more details in Appendix B). This approach enabled the generation of over 620,000 legal queries from 140,000 passages in the VNLAWQC dataset. An example of a generated query and its corresponding passage is shown in Table 3. Finally, we merge VNLAWQC and VNSYNLAWQC with the ZaloLegal2021 (Zalo AI Team, 2021) and employ deduplication to prevent data leakage and improve data quality.

Cross-Lingual Dataset using Translation: To facilitate the CLIR scenario, we leverage translation models to produce Vietnamese and English versions of both queries and documents. We integrate multiple datasets, as presented in Table 1, encompassing general and legal domain knowledge in both languages. For the Vietnamese datasets, we employ the VinAI Translate (Nguyen et al., 2022) to generate English versions, while Google Translate is used to translate the English datasets into Vietnamese. Both models have demonstrated state-of-the-art performance on Vietnamese-English translation benchmarks, such as PhoMT (Doan et al., 2021), highlighting their suitability and effectiveness for our task.

To ensure translation quality, we use back-translation and evaluate the similarity between the original text and its back-translated version with Jaccard similarity (Jaccard, 1912; Tanimoto, 1958). Translation pairs that have a score below the predetermined threshold of **0.5** are then discarded. Additionally, to further assess the quality, we manually reviewed 100 randomly selected samples and verified that they meet satisfactory standards. This multi-step process guarantees that the translations are of high quality for constructing the cross-lingual dataset.

Hard Negatives Mining: To further enhance the training process, we provide hard negatives (e.g., documents) for each example (e.g., query), which offer more informative negative samples and potentially improve training convergence (Xiong et al., 2021). Specifically, we utilize a retrieval model like BGE-M3 (Chen et al., 2024) to identify the most similar documents and adopt a threshold score to guarantee the selection of true negative samples. We gathered these samples and added additional random contexts, if necessary, to create five negative candidates for each query.

3.2 Cross-Lingual Retrieval Model

Embedding Backbone: We choose the pre-trained BGE-M3 (Chen et al., 2024), which can support three retrieval modes: Dense, Lexical, and Multi-Vector, as the backbone model. It supports a long context window of up to 8,192 tokens and is pre-trained in multiple languages, including Vietnamese, which is beneficial for retrieving lengthy legal documents and handling cross-lingual tasks involving Vietnamese. The pre-trained BGE-M3 model is also used as the baseline for evaluating the improvements made by our method.

Auxiliary Loss Function: The original BGE-M3 embedding model employs two primary loss functions: $\mathcal{L}^{InfoNCE}$, an InfoNCE loss (Oord et al., 2018) that controls the alignment between queries and both positive and negative passages, and a self-knowledge distillation loss $\mathcal{L}^{distill}$, which allows the multiple retrieval modes to be jointly learned and mutually reinforced. In cross-lingual scenarios, queries tend to be short and ambiguous. To address this, we propose a loss function that improves alignment between each query and its translated version.

$$\mathcal{L}^{aux} = -\log \frac{\exp(s(q, \bar{q})/\tau)}{\sum_{a \in Q} \exp(s(q, \bar{a})/\tau)}$$

where q is a query, $\bar{q}$ is its translated version of q , Q is the set of queries in a batch and τ is the temperature hyperparameter. Consequently, we combine these loss functions to train our model: $\mathcal{L} = \mathcal{L}^{InfoNCE} + \mathcal{L}^{distill} + \mathcal{L}^{aux}$.

Symmetrical Training: Currently, retrieval models are trained to minimize the distance between a query and its corresponding relevant documents. We extend this by introducing Symmetrical Training to learn relationships between similar queries and documents across languages. In this approach, a document or query in one language is treated as relevant to its translated version. Given two versions of a document or query, S_A in language A and S_B in language B, we then consider S_A and S_B as a valid training pair. The model is finetuned to retrieve the translated version of a query or document with a fixed probability, p_{sym}, alongside the standard query-document retrieval task. Hard negatives for these symmetrical pairs are mined similarly to unsymmetrical ones.

4 Experiments and Results

Experiment Setup: We trained the models for 4 epochs using the AdamW optimizer (Loshchilov and Hutter, 2019) with a base learning rate of 2×10^{-5}. A cosine learning rate scheduler with the warm-up ratio set to 5% of the total training steps was applied. The temperature τ was set to 0.05. Besides, we employed smart batching (Ge et al., 2021) to group samples with similar sequence lengths. For symmetrical training, the sampling rate p_{sym} was set to 0.3. The trained models are subsequently evaluated on the evaluation sets from VNLAWQC and ZaloLegal2021 for the legal domain, as well as ZaloWikipediaQA for the general knowledge domain. We conduct evaluations using various training datasets, retrieval modes, and loss functions.

Evaluation Metric: Following prior work in document retrieval (Karpukhin et al., 2020; Wang et al., 2022; Neelakantan et al., 2022; Dai and Callan, 2020; Khattab and Zaharia, 2020), we leverage four metrics Recall@k, MRR@k, MAP@k and nDCG@k for evaluation. Specifically, we use $k = 10$ and calculate the average of these four metrics for performance comparison. We observe that the average scores exhibit a strong correlation with individual metrics, making them a suitable representation of overall performance. Detailed results for each specific metric are provided in Tables 6 and 7.

Experimental Results: Table 2 presents the results of the baseline models and our proposed models across two cross-lingual scenarios: Vietnamese-English, where queries are in English and documents are in Vietnamese, and English-Vietnamese, where the roles are reversed. Additional results on monolingual scenarios and detailed metrics for cross-lingual tasks are provided in Appendix C.1 and C.2, respectively. In summary, our proposed datasets and methods enhance the performance of the multilingual embedding backbone (i.e No training), achieving scores that rank among the highest across all evaluation sets. Besides, the retrieval mode with reranking consistently outperforms dense retrieval alone. This improvement is evident due to the additional ranking stage, which enhances the selection of relevant documents, although it also incurs extra costs.

- **Effectiveness of Cross-lingual Data:** The results show that the BGE-M3 model fine-tuned on cross-lingual data significantly outperforms both the original model and the one fine-tuned on Vietnamese data. Specifically, we observe an improvement of over 10% in legal document retrieval and more than 3% in the general domain. This observation further highlights the quality of our construction pipeline with translation.
- **Effectiveness of Synthetic Data:** The inclusion of VNSYNLAWQC during training generally enhances the performance of all models across the

Training	Synthetic	Retrieval	Vietnamese-English			English-Vietnamese		
Approach	Augmentation	Mode	VNLawQC	ZaloLegal2021	ZaloWikipediaQA	VNLawQC	ZaloLegal2021	ZaloWikipediaQA
No training	✗	D	42.99	51.88	64.84	39.46	48.26	62.31
		$D+R$	44.92	53.63	66.18	40.81	49.16	63.45
Vietnamese	✗	D	54.14	62.26	65.96	47.38	55.60	61.01
		$D+R$	56.78	65.14	71.14	50.48	59.03	67.48
	✓	D	56.18	62.42	64.60	48.90	53.62	60.05
		$D+R$	58.32	65.72	70.25	53.38	57.26	66.57
Cross-lingual	✗	D	68.02	75.32	67.90	65.39	71.33	65.06
		$D+R$	70.21	77.57	72.90	68.04	74.33	70.88
	✓	D	69.44	78.74	68.61	66.89	75.42	66.12
		$D+R$	71.58	**80.55**	73.88	68.95	78.41	71.54
Cross-lingual +aux_loss_function	✓	D	68.60	75.49	69.56	66.18	73.18	66.39
		$D+R$	71.46	78.62	**74.18**	69.53	76.18	**71.78**
Cross-lingual +sym_training	✓	D	69.75	76.33	65.64	67.66	75.97	62.15
		$D+R$	**71.71**	79.53	68.75	**69.91**	**79.75**	66.73

Table 2: Performance of BGE-M3 in CLIR scenario using different training methods, datasets, and retrieval mode across three evaluation sets in legal and general knowledge domains. In the training approach, *Cross-lingual* refers to the use of datasets in both language versions, while *aux_loss_function* and *sym_training* indicate the loss function and Symmetrical Training described in Section 3.2. *Synthetic Augmentation* refers to the use of VNSynLawQC to augment the training data during the training process. In retrieval modes, D represents dense retrieval results, while $D + R$ represents the results when a reranking stage is incorporated for the retrieved documents. **Green scores** indicate the highest score, while Gray scores represent the second highest.

evaluation datasets. In particular, an improvement of nearly 3% is observed in the ZaloLegal2021 dataset for the Vietnamese-English scenario, achieving the highest performance with a score of 80.55%. Similarly, all evaluation sets showed improvements when using synthetic data in the English-Vietnamese scenario.

- **Effectiveness of Auxiliary Loss and Symmetrical Training:** The implementation of auxiliary loss functions and symmetrical training yields varying results depending on the dataset domain. While models with symmetrical training demonstrate significant performance in legal retrieval, models trained with auxiliary loss achieve the highest performance in the general knowledge domain. These results align with our motivation for employing auxiliary loss, as queries in the general domain tend to be short and ambiguous.

5 Conclusion

In summary, we introduce a novel dataset for cross-lingual information retrieval (CLIR) between Vietnamese and English, covering both general knowledge and the legal domain. Additionally, we develop a CLIR model by finetuning cross-lingual and synthetic data while proposing an auxiliary loss function and training strategy to enhance performance. Our contributions provide valuable resources and methods for advancing cross-lingual retrieval in specialized fields.

6 Limitations

The proposed dataset, reliant on translation techniques, may be prone to translation errors and may not fully reflect real-world data patterns. To mitigate this issue, we have made efforts by implementing quality control measures during the generation process to ensure the quality and naturalness of the translations. However, we recommend that mining real-world data or human intervention is crucial for effectively addressing this issue.

In this study, our experiments utilize a single backbone model, which may raise concerns regarding the versatility and adaptability of the proposed methodologies. The backbone model employed in our study, BGE-M3, has already demonstrated state-of-the-art performance across multiple document retrieval benchmarks. As a result, the enhancements observed in this model can well prove the effectiveness of our methodologies. In future work, we aim to extend our techniques to a broader array of models to gain deeper insights into their robustness and adaptability, thereby advancing cross-lingual information retrieval research.

Acknowledgements

This work was funded by the NAVER Corporation within the framework of collaboration with the International Research Center for Artificial Intelligence (BKAI), School of Information and Communication Technology, Hanoi University of Science and Technology. Nguyen Doan Hieu was funded by the Master, PhD Scholarship Program of Vingroup Innovation Foundation (VINIF), code VINIF.2022.ThS.BK.09.

References

Hamed Bonab, Sheikh Muhammad Sarwar, and James Allan. 2020. Training effective neural clir by bridging the translation gap. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 9–18.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Zhuyun Dai and Jamie Callan. 2020. Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1533–1536.

Long Doan, Linh The Nguyen, Nguyen Luong Tran, Thai Hoang, and Dat Quoc Nguyen. 2021. Phomt: A high-quality and large-scale benchmark dataset for vietnamese-english machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 4495–4503.

Qian Dong, Yiding Liu, Qingyao Ai, Haitao Li, Shuaiqiang Wang, Yiqun Liu, Dawei Yin, and Shaoping Ma. 2023. I3 retriever: incorporating implicit interaction in pre-trained language models for passage retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 441–451.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: revisit exact lexical match in information retrieval with contextualized inverted list. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, pages 3030–3042.

Zhenhao Ge, Lakshmish Kaushik, Masanori Omote, and Saket Kumar. 2021. Speed up training with variable length inputs by efficient batching strategies. In *Interspeech*, pages 156–160.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1048–1056.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. 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*, pages 39–48.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Jiapeng Liu, Xiao Zhang, Dan Goldwasser, and Xiao Wang. 2020. Cross-lingual document retrieval with smooth learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3616–3629.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of ICLR 2019*.

Antoine Louis, Vageesh Kumar Saxena, G. van Dijck, and Gerasimos Spanakis. 2024. Colbert-xm: A modular multi-vector representation model for zero-shot multilingual information retrieval. *ArXiv*, abs/2402.15059.

Antoine Louis and Gerasimos Spanakis. 2022. A statutory article retrieval dataset in french. In *Proceedings of ACL 2022*, pages 6789–6803.

Shengmei Luo, Guangyan Zhang, Chengwen Wu, Samee U Khan, and Keqin Li. 2015. Boafft: Distributed deduplication for big data storage in the cloud. *IEEE transactions on cloud computing*, 8(4):1199–1211.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. In *Findings of EMNLP 2020*, volume EMNLP 2020, pages 1037–1042.

Ha-Thanh Nguyen, Manh-Kien Phi, Xuan-Bach Ngo, Vu Tran, Le-Minh Nguyen, and Minh-Phuong Tu. 2024. Attentive deep neural networks for legal document retrieval. *Artificial Intelligence and Law*, 32(1):57–86.

Thien Hai Nguyen, Tuan Duy H Nguyen, Duy Phung, Duy Tran Cong Nguyen, Hieu Minh Tran, Manh Luong, Tin Duy Vo, Hung Hai Bui, Dinh Phung, and Dat Quoc Nguyen. 2022. A vietnamese-english neural machine translation system. In *Annual Conference of the International Speech Communication Association (was Eurospeech) 2022*, pages 5543–5544. International Speech Communication Association (ISCA).

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Yifu Qiu, Hongyu Li, Yingqi Qu, Ying Chen, Qiaoqiao She, Jing Liu, Hua Wu, and Haifeng Wang. 2022. Dureader-retrieval: A large-scale chinese benchmark for passage retrieval from web search engine. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5326–5338.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of ACL 2018*, pages 784–789.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797.

Carlo Sansone and Giancarlo Sperlí. 2022. Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106:101967.

Weihang Su, Yiran Hu, Anzhe Xie, Qingyao Ai, Quezi Bing, Ning Zheng, Yun Liu, Weixing Shen, and Yiqun Liu. 2024. Stard: A chinese statute retrieval dataset derived from real-life queries by non-professionals. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10658–10671.

Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. 2019. Legal document retrieval using document vector embeddings and deep learning. In *Intelligent Computing: Proceedings of the 2018 Computing Conference, Volume 2*, pages 160–175. Springer.

Taffee T Tanimoto. 1958. Elementary mathematical theory of classification and prediction.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. Vncorenlp: A vietnamese natural language processing toolkit. In *Proceedings of NAACL-HLT 2018*, pages 56–60.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Zalo AI Team. 2019. Zalo challenge dataset. Zalo AI Challenge 2019, `https://challenge.zalo.ai`. Accessed: 2025-02-08.

Zalo AI Team. 2021. Zalo challenge dataset. Zalo AI Challenge 2021, `https://challenge.zalo.ai`. Accessed: 2025-02-08.

Eric Zhu and Vadim Markovtsev. 2017. ekzhu/datasketch: First stable release. Accessed: 2025-02-08.

Appendix

A Additional Details on Data Construction

B Synthetic Data Generation

Prompt A. Synthetic Data Template

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:
<document>
<domain>{DOC_DOMAIN}</domain>
<source>{DOC_SOURCE}</source>
<content>{DOC_GROUNDED_CONTENT}</content>
</document>
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 thể 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.
<example>
<description>Bad questions refer to the legal text directly</description>
<bad_question> Thông tư này quy định những nguyên tắc gì trong việc thi hành án tử hình bằng hình thức tiêm thuốc độc?</bad_question>
<good_question>Pháp luật quy định những nguyên tắc gì trong việc thi hành án tử hình bằng hình thức tiêm thuốc độc?</good_question>
<best_question>Thông tư do Bộ Công an ban hành quy định những nguyên tắc gì trong việc thi hành án tử hình bằng hình thức tiêm thuốc độc?</best_question>
</example>
<example>
<description>Bad questions does not include enough context or detail</description>
<bad_question> Theo quy định, người được khám giám định không đồng ý với kết quả khám giám định phúc quyết của Hội đồng Giám định Y khoa cấp Trung ương thì sẽ được xử lý như thế nào?</bad_question>
<good_question>Nếu người bị phơi nhiễm chất độc hóa học trong kháng chiến không đồng ý với kết quả giám định của Hội đồng GĐYK cấp Trung ương, họ có thể làm gì để được xem xét lại? </good_question>
</example>
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:

	Vietnamese	English
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.	Section 1. INSPECTION PREPARATION, Chapter II. PROCEDURES AND PROCESSES FOR CONDUCTING INSPECTIONS ACCORDING TO THE INSPECTION PLAN, Circular 36/2016/TT-NHNN stipulating the procedures and processes for specialized banking inspections, issued by the Governor of the State Bank of Vietnam.
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.	5. The Head of the Inspection team organizes a meeting with the Inspection team to disseminate the approved inspection plan and assign tasks to the Inspection groups, Inspection units, and members of the Inspection team; discuss and decide on the methods and organization of the inspection process; and coordinate among members of the inspection team and related agencies or units during the inspection. If necessary, the person who issued the inspection decision or an authorized representative may attend the meeting to emphasize the purpose, requirements, and content of the inspection, as well as the responsibilities of the Inspection team. Task assignments for the Inspection groups, units, and team members must be documented in writing. 6. The Inspection group leaders, Inspection unit leaders, and members of the Inspection team shall develop plans to carry out their assigned tasks and report to the Head of the Inspection team before conducting the inspection at the credit institution.
Aspect 1	Trách nhiệm của Trưởng đoàn thanh tra trong việc tổ chức và phân công nhiệm vụ	Responsibilities of the Head of the Inspection Team in organizing and assigning tasks
Query 1	Ngân hàng Nhà nước quy định Trưởng đoàn thanh tra phải làm gì để chuẩn bị cho cuộc thanh tra?	What does the State Bank require the Head of the Inspection Team to do to prepare for the inspection?
Aspect 2	Quy trình xây dựng và báo cáo kế hoạch thực hiện nhiệm vụ của các Tổ thanh tra, Nhóm thanh tra	The process of developing and reporting task execution plans by the Inspection groups and Inspection units
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?	When assigned tasks, what must the Inspection groups and Inspection units do to prepare for the inspection?

Table 3: Example of a generated query-passage pair for the domain "Tiền tệ-Ngân hàng" (Currency-Banking)

B.1 Generate synthetic queries

For generating synthetic queries, we utilized the open-source large language model Meta Llama 3 (Dubey et al., 2024) to generate queries based on aspects identified within legal text passages. This process involved extracting key aspects from the texts and formulating corresponding queries. We selected Llama-3-70B for its strong capabilities and performance. Additionally, Llama 3 is believed to include a portion of synthetic data in its training corpus. Upon release, it outperformed many other models with a similar parameter count, demonstrating notable proficiency across multiple languages, including Vietnamese, aligning well with our requirements.

A significant challenge in using LLMs for query generation is maintaining both the diversity and relevance of their outputs. We experimented with different prompt techniques to achieve this balance. One approach instructed the model to generate questions directly from the passage without first identifying different aspects. This method often resulted in less diverse and sometimes irrelevant queries, as the model tended to focus on the most prominent information in the passage, neglecting other potential aspects.

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. The prompt template used for generating these synthetic queries is illustrated in prompt

B. Applying this method, we generated over 620,000 legal queries from 140,000 passages in VNLAWQC dataset. An example of a generated query and its corresponding passage is shown in Table 3.

B.2 Filter low-quality queries

After generating the synthetic data, we removed low-quality queries that explicitly referred to the input passage or were only shallowly relevant to the passage content. In particular, we employed the BGE-M3 dense retriever (Chen et al., 2024), which demonstrated strong zero-shot 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*" (*this regulation*) or "*thông tư này*" (*this circular*). This process resulted in the final VNSYNLAWQC dataset, which contains over 500,000 high-quality queries.

C Additional Experimental Results

In this section, we present additional results for both mono-lingual (Section C.1) and cross-lingual (Section C.2) settings. Additionally, we explore different reranking modes, as discussed in Section 4. For reranking, we employ the multi-vector mode, which incurs minimal overhead since it is trained concurrently with dense retrieval. Only the top 100 passages from dense retrieval are reranked to reduce computational cost. Reranking times were measured on Kaggle's T4 and an RTX3090: cross-encoder reranking (BGE-reranker-v2-m3) took 7.15s/query (T4) and 1.33s/query (RTX3090), while our multi-vector mode took 6.41s/query (T4) and 1.05s/query (RTX3090).

C.1 Mono-lingual Retrieval Results

Training Approach	Synthetic Augmentation	Retrieval Mode	**VNLAWQC**				**ZaloLegal2021**				**ZaloWikipediaQA**			
			R@10	**MRR@10**	**MAP@10**	**nDCG@10**	**R@10**	**MRR@10**	**MAP@10**	**nDCG@10**	**R@10**	**MRR@10**	**MAP@10**	**nDCG@10**
No training	✗	D	65.09	43.73	42.06	48.07	73.07	49.67	49.39	55.13	85.25	65.61	63.00	69.31
		$D+R$	65.88	45.22	43.44	49.33	75.86	52.96	52.69	58.31	86.95	67.35	64.99	71.17
Vietnamese	✗	D	73.31	51.42	49.40	55.73	81.95	58.04	57.68	63.64	82.45	64.45	61.54	67.56
		$D+R$	76.14	55.06	53.01	59.18	84.69	63.22	62.97	68.29	87.14	69.25	66.89	72.65
	✓	D	73.67	52.11	50.07	56.33	82.89	60.39	60.16	65.74	82.10	64.08	61.13	67.17
		$D+R$	76.38	55.53	53.42	59.57	83.83	63.49	63.26	68.35	86.90	69.24	66.77	72.52
Cross-lingual	✗	D	80.93	60.44	58.36	64.43	86.22	67.65	67.43	72.07	83.20	64.42	61.54	67.72
		$D+R$	82.74	63.39	61.22	67.06	88.72	70.85	70.62	75.08	87.33	69.42	66.94	72.75
	✓	D	81.59	61.19	59.08	65.14	88.93	69.58	69.20	74.10	83.24	64.94	62.04	68.12
		$D+R$	83.05	63.67	61.49	67.36	**89.43**	**72.28**	**71.98**	**76.33**	87.61	70.06	67.46	73.25
Cross-lingual +aux_loss_function	✓	D	81.19	60.60	58.47	64.58	85.21	65.82	65.54	70.41	84.50	65.96	63.13	69.24
		$D+R$	83.03	63.98	61.78	67.56	88.54	69.14	68.83	73.74	**88.32**	**70.55**	**68.11**	**73.87**
Cross-lingual +sym_loss	✓	D	81.74	62.03	59.96	65.84	86.85	65.97	65.69	70.91	79.69	60.20	57.29	63.67
		$D+R$	**83.10**	**64.32**	**62.14**	**67.84**	87.01	69.16	68.95	73.41	82.68	63.30	60.64	66.90

Table 4: English-English retrieval results on different datasets. Both the queries and the documents are in English.

Table 4 presents the performance of our cross-lingual models in the English-English retrieval setting. All cross-lingual models significantly outperform the baseline on both legal datasets. Notably, the cross-lingual model with symmetrical training achieves the highest R@10 score of 83.10% on the VNLAWQC dataset, while the base cross-lingual model attains the highest R@10 score of 89.43% on the ZaloLegal2021 dataset. In contrast, for the ZaloWikipediaQA dataset, although there is a slight decline in dense retrieval performance, incorporating reranking and the auxiliary loss function boosts the cross-lingual model to an optimal R@10 of 88.32%.

However, we noticed that on the two legal datasets, despite having higher performance compared to the baseline model, the performance is lower than in the Vietnamese-English setting. We hypothesize that this issue arises from errors propagated during the translation process. While the documents typically contain multiple sentences and are sufficiently lengthy to provide contextual information, the queries are short and consist of only a single sentence, which may lead to translation inaccuracies due to the lack of contextual cues.

We further evaluated our cross-lingual models in the Vietnamese-only retrieval setting. As presented in Table 5, despite being trained on cross-lingual data, these models perform comparably to the Vietnamese model, which was trained exclusively on Vietnamese data. On both legal datasets, the cross-lingual models surpass the baseline and achieve competitive results. For the VNLAWQC dataset, the cross-lingual model augmented with the auxiliary loss function attains an R@10 of 86.5%, which is only marginally

Training Approach	Synthetic Augmentation	Retrieval Mode	VNLawQC				ZaloLegal2021				ZaloWikipediaQA			
			R@10	**MRR@10**	**MAP@10**	**nDCG@10**	**R@10**	**MRR@10**	**MAP@10**	**nDCG@10**	**R@10**	**MRR@10**	**MAP@10**	**nDCG@10**
No training	✗	D	73.93	52.93	50.94	57.05	81.46	59.08	58.76	64.31	94.26	76.82	74.49	80.18
		$D+R$	75.22	54.68	52.62	58.66	82.47	61.73	61.43	66.61	**95.54**	78.56	76.45	81.91
Vietnamese	✗	D	85.53	66.04	63.92	69.79	91.38	73.90	73.69	78.11	92.11	74.65	71.96	77.76
		$D+R$	86.76	68.67	66.52	72.06	**94.06**	77.04	76.78	81.08	95.23	**78.93**	**76.82**	**82.09**
	✓	D	86.33	67.98	65.72	71.37	91.67	75.07	74.80	79.02	91.25	73.74	70.96	76.83
		$D+R$	**87.33**	**70.66**	**68.35**	**73.62**	93.54	**79.75**	**79.39**	**82.99**	94.83	78.87	76.69	81.91
Cross-lingual	✗	D	84.45	65.53	63.24	69.03	90.05	70.79	70.58	75.38	89.26	71.54	68.71	74.64
		$D+R$	86.10	68.36	66.07	71.58	91.95	73.35	73.09	77.79	93.69	76.97	74.64	80.11
	✓	D	84.93	65.89	63.69	69.48	90.86	75.12	74.79	78.82	89.54	72.05	69.13	75.05
		$D+R$	86.30	68.65	66.41	71.88	92.60	77.06	76.76	80.78	93.84	77.38	75.15	80.53
Cross-lingual +aux_loss_function	✓	D	84.95	65.63	63.43	69.27	91.28	72.36	72.02	76.86	90.75	73.12	70.36	76.23
		$D+R$	86.50	69.24	66.98	72.38	92.68	75.14	74.84	79.33	94.63	78.13	76.03	81.35
Cross-lingual +sym_loss	✓	D	84.73	66.59	64.48	70.00	91.64	75.09	74.82	79.01	85.99	66.50	63.42	69.90
		$D+R$	85.68	69.02	66.76	72.00	92.58	76.82	76.56	80.56	88.03	68.74	66.07	72.33

Table 5: Vietnamese-Vietnamese retrieval results on different datasets. Both the queries and the documents are in Vietnamese.

lower than the Vietnamese model's score of 87.33% under the same dense + re-ranking pipeline with synthetic augmentation. Similarly, on the ZaloLegal2021 dataset, it also achieves an R@10 of 92.68%, closely aligning with the Vietnamese model's top score of 94.06%. Although performance declines on the ZaloWikipediaQA dataset, the use of reranking and auxiliary loss still helps the cross-lingual model achieve an R@10 of 94.63%, outperforming other configurations. The use of synthetic augmentation generally leads to performance improvements across all training approaches, except for the Vietnamese model on the ZaloWikipediaQA dataset, where the gains are less pronounced.

C.2 Cross-lingual Retrieval Results

Training Approach	Synthetic Augmentation	Retrieval Mode	VNLawQC				ZaloLegal2021				ZaloWikipediaQA			
			R@10	**MRR@10**	**MAP@10**	**nDCG@10**	**R@10**	**MRR@10**	**MAP@10**	**nDCG@10**	**R@10**	**MRR@10**	**MAP@10**	**nDCG@10**
No training	✗	D	54.65	34.51	33.06	35.61	66.77	41.35	41.15	43.75	78.24	57.54	54.99	58.47
		$D+R$	56.50	35.63	34.20	36.91	66.17	42.52	42.27	45.68	79.05	58.64	56.17	59.92
Vietnamese	✗	D	63.85	42.03	40.34	43.30	72.89	48.91	48.65	51.95	76.55	56.54	53.76	57.20
		$D+R$	66.91	45.09	43.38	46.52	76.33	52.31	52.06	55.41	82.34	63.00	60.50	64.09
	✓	D	65.26	43.60	41.86	44.86	69.79	47.20	46.98	50.49	75.34	55.72	52.94	56.22
		$D+R$	69.34	48.25	46.42	49.52	73.54	51.04	50.83	53.64	81.29	62.13	59.61	63.24
Cross-lingual	✗	D	81.15	60.34	58.31	61.74	84.77	65.90	65.74	68.91	80.09	60.61	57.90	61.65
		$D+R$	82.64	63.42	61.33	64.77	87.94	68.99	68.73	71.64	85.36	66.44	64.07	67.66
	✓	D	81.90	62.10	59.99	63.58	88.62	70.24	69.86	72.96	80.93	61.82	59.06	62.65
		$D+R$	83.24	64.42	62.27	65.87	90.94	73.50	73.21	76.00	**86.02**	67.09	64.68	68.37
Cross-lingual +aux_loss_function	✓	D	81.79	61.18	59.10	62.64	87.63	67.32	67.07	70.70	80.96	62.14	59.40	63.06
		$D+R$	**83.77**	65.11	62.91	66.33	90.13	70.66	70.34	73.59	85.96	**67.28**	**65.07**	**68.81**
Cross-lingual +sym_loss	✓	D	81.96	63.11	61.08	64.49	88.88	70.98	70.70	73.33	77.95	57.50	54.76	58.41
		$D+R$	83.51	**65.64**	**63.56**	**66.95**	**91.17**	**75.27**	**74.98**	**77.59**	81.36	62.28	59.76	63.52

Table 6: English-Vietnamese retrieval results on different datasets. The queries are in Vietnamese and the documents are in English.

Training Approach	Synthetic Augmentation	Retrieval Mode	VNLawQC				ZaloLegal2021				ZaloWikipediaQA			
			R@10	**MRR@10**	**MAP@10**	**nDCG@10**	**R@10**	**MRR@10**	**MAP@10**	**nDCG@10**	**R@10**	**MRR@10**	**MAP@10**	**nDCG@10**
No training	✗	D	58.25	36.91	35.47	41.33	69.79	43.95	43.72	50.05	80.00	59.43	56.68	63.26
		$D+R$	60.44	38.76	37.23	43.23	70.73	46.11	45.83	51.86	81.39	60.66	58.05	64.62
Vietnamese	✗	D	70.29	47.91	45.99	52.37	79.30	54.69	54.51	60.53	80.44	61.07	57.95	64.38
		$D+R$	72.95	50.58	48.59	55.01	81.30	57.96	57.78	63.53	85.25	66.20	63.45	69.65
	✓	D	71.42	50.37	48.42	54.51	81.72	53.82	53.64	60.48	79.60	59.33	56.46	62.99
		$D+R$	73.36	52.61	50.62	56.67	82.66	58.17	57.99	64.05	84.34	65.31	62.59	68.77
Cross-lingual	✗	D	82.70	62.59	60.38	66.42	87.55	69.94	69.68	74.12	82.22	62.95	60.07	66.37
		$D+R$	84.14	65.13	62.89	68.69	89.61	72.27	72.01	76.41	86.48	68.16	65.48	71.48
	✓	D	84.06	63.97	61.86	67.85	90.55	73.59	73.23	77.57	83.16	63.61	60.63	67.05
		$D+R$	85.27	66.57	64.38	70.08	**92.34**	**75.38**	**75.09**	**79.40**	87.32	69.23	66.50	72.46
Cross-lingual +aux_loss_function	✓	D	83.65	62.97	60.82	66.97	88.85	69.59	69.32	74.19	83.67	64.67	61.84	68.07
		$D+R$	**85.32**	66.42	64.17	69.94	91.48	73.00	72.62	77.36	**87.63**	**69.40**	**66.88**	**72.80**
Cross-lingual +sym_training	✓	D	83.37	64.71	62.63	68.27	89.19	70.64	70.40	75.07	80.40	60.59	57.52	64.03
		$D+R$	84.54	**67.11**	**64.90**	**70.30**	90.62	74.65	74.40	78.46	83.48	63.48	60.83	67.20

Table 7: Vietnamese-English retrieval results on different datasets. The queries are in Enlish and the documents are in Vietnamese.

We finally evaluated our models on English-Vietnamese and Vietnamese-English cross-lingual retrieval tasks, as presented in Tables 6 and 7. The results indicate that for both retrieval directions, our cross-lingual models consistently outperform the baseline and the Vietnamese version, achieving the highest performance across all metrics and datasets, including the ZaloWikipediaQA dataset. This superior performance suggests a robust understanding of the semantic relationships between Vietnamese and English content.

For the English-Vietnamese retrieval task, the cross-lingual model with an auxiliary loss function achieves an R@10 of 83.77% on the VNLawQC dataset, which is 27% higher than the baseline and 14% higher than the Vietnamese model. Similarly, in the ZaloLegal2021 dataset, the cross-lingual model with symmetrical training achieves an R@10 of 91.17%, which is 24% higher than the baseline and 15% higher than the Vietnamese model. On the ZaloWikipediaQA dataset, the cross-lingual model records an R@10 of 86.03%, outperforming the baseline by 7% and the Vietnamese model by 4%.

For the Vietnamese-English retrieval task, the cross-lingual models achieve even higher results. On the VNLawQC dataset, the best cross-lingual model attains an R@10 of 85.32%, which is 25% higher than the baseline and 12% higher than the Vietnamese model. In the ZaloLegal2021 dataset, the cross-lingual model achieves an R@10 of 92.34%, reflecting a 22% increase over the baseline and a 10% improvement over the Vietnamese model. For the ZaloWikipediaQA dataset, the cross-lingual model reaches an R@10 of 87.63%, surpassing the baseline by 6% and the Vietnamese model by 3%. These findings underscore the effectiveness of our cross-lingual models, particularly when combined with the Auxiliary Loss Function and Symmetrical Training strategies.

Furthermore, synthetic augmentation results in an average performance improvement of 1% across all datasets. Notably, an improvement of 3% is observed in the ZaloLegal2021 dataset for both the English-Vietnamese and the Vietnamese-English scenario, highlighting its positive impact on retrieval effectiveness.