ViDRILL: A Multi-Stage Retrieval Framework for Vietnamese Legal Document Search

Dien X. Tran*[1], Tai D. Truong[1], Kien C. Nguyen*[1]

[1]Industrial University of Ho Chi Minh City, Vietnam

* Corresponding authors.
22650601.dien@student.iuh.edu.vn
nguyenchikien@iuh.edu.vn

Abstract

Legal information retrieval in Vietnamese remains underexplored despite the growing demand for intelligent legal NLP systems. This paper presents ViDRILL, a multi-stage retrieval framework designed for the VLSP 2025 DRiLL shared task on Vietnamese legal document retrieval. Our approach integrates sparse and dense retrieval with semantic re-ranking: BM25 provides strong lexical baselines, multilingual embeddings (E5-Instruct, GTE) capture semantic relevance, and a cross-encoder re-ranker (BGE-rerank) refines candidate rankings. To enhance training effectiveness, we introduce a dual-level chunking strategy and a hard negative sampling mechanism guided by pretrained models. Experimental results on the official benchmark demonstrate that ViDRILL achieves top-5 performance, highlighting the effectiveness of combining lexical precision, semantic retrieval, and cross-encoder re-ranking for complex legal texts.

1 Introduction

Access to legal information plays a central role in supporting decision-making for lawyers, judges, and citizens. However, retrieving the right legal documents from large collections is a non-trivial task. Legal texts are typically lengthy, formal, and densely interconnected, with clauses that reference multiple articles or laws. Queries in this domain are expressed in natural language, which introduces additional challenges of ambiguity, paraphrasing, and mismatch between layman expressions and formal legal terminology. These factors make conventional retrieval techniques insufficient for the legal domain (Nguyen et al., 2024).

In the Vietnamese context, research on legal information retrieval (LIR) is still at an early stage. While progress has been made in general-purpose Vietnamese NLP tasks such as machine translation, summarization, and question answering (QA), domain-specific retrieval systems remain scarce. This gap is particularly critical given the increasing digitization of Vietnamese legal documents and the growing demand for intelligent legal NLP tools. Recent studies have begun to address this challenge: (Tien et al., 2024) generated synthetic queries to fine-tune retrieval models, while (Nguyen et al., 2025) demonstrated that optimized data processing, tailored loss functions, and semi-hard negative sampling are pivotal for building robust retrieval-augmented systems. These advances highlight the importance of moving beyond purely lexical methods toward hybrid and multi-stage approaches for Vietnamese legal texts.

To further bridge this gap, the VLSP 2025 DRiLL (Document Retrieval in Legal Language) shared task was introduced as a benchmark for evaluating retrieval systems on Vietnamese legal corpora (Vuong et al., 2025). The task emphasizes not only lexical matching but also deeper semantic understanding, encouraging systems that balance recall and precision across complex queries.

In this work, we present ViDRILL, our multi-stage retrieval framework for the DRiLL task. The system integrates BM25 for strong lexical baselines (Robertson and Zaragoza, 2009), dense multilingual encoders (E5-Instruct, GTE, BGE-M3) (Wang et al., 2024; Zhang et al., 2024; Chen et al., 2024) for semantic retrieval, and a cross-encoder re-ranker (BGE-reranker-v2-m3) for fine-grained ranking (Nogueira and Cho, 2019). Beyond retrieval accuracy, we propose a dual-level chunking strategy and a two-stage hard negative mining mechanism, tailored specifically to long and structured Vietnamese legal texts. Through this design, our contributions aim to advance the development of robust retrieval frameworks for Vietnamese legal NLP and provide practical insights for handling complex legal search scenarios.

2 Related Works

Research on legal information retrieval (LIR) has progressed from keyword-based methods to neural architectures. While lexical approaches focus on exact matching, recent work emphasizes semantic representations and multi-stage pipelines. In Vietnamese, studies remain limited, though new datasets and shared tasks are driving advances. We review related work in two directions: (i) lexical vs. dense retrieval and (ii) re-ranking and multi-stage pipelines for legal NLP.

2.1 Legal Information Retrieval: Lexical vs. Dense Retrieval

Early legal IR methods relied on lexical matching such as TF–IDF and BM25, which remain strong baselines due to robust term–document exact matching and efficient indexing (Robertson and Zaragoza, 2009). Nevertheless, lexical approaches often fail at capturing synonymy, paraphrase, and long, compositional legal provisions. Recently, neural dense retrieval approaches such as legal case-specific encoders (e.g., CaseEncoder with Biased Circle Loss (Ma et al., 2023) and DELTA using structural word alignment (Li et al., 2024)) have emerged, showing notable improvements in semantic matching for legal case retrieval. A recent survey also highlights growing interest and performance gains in legal case retrieval using domain-aware dense encoders (Feng et al., 2024). In addition, specialized approaches for statute law (e.g., combining BM25 with Longformer for long input handling) have proven effective in COLIEE tasks (Nguyen et al., 2022). Unlike these single-strategy approaches, ViDRILL uniquely integrates lexical precision with dense semantic recall into a unified framework tailored for Vietnamese legal documents.

2.2 Re-ranking and Multi-stage Pipelines in Legal NLP

Two-stage pipelines lexical/dense first-stage retrieval followed by cross-encoder re-ranking are now standard practice in legal IR, aiming to balance efficiency and accuracy effectively. For instance, the COLIEE 2022 entry LeiBi combined tuned lexical models with cluster-driven BERT-based re-ranking, yielding notable improvements in case law retrieval (Askari et al., 2022). Similarly, the DoSSIER approach (COLIEE 2021) leveraged dense retrieval at the paragraph level and summarization-based BERT re-ranking to tackle long legal cases (Althammer et al., 2021). A more recent architecture, TraceRetriever, segments documents via rhetorical roles and employs BM25, vector retrieval, and cross-encoder models fused with Reciprocal Rank Fusion, achieving state-of-the-art results on IL-PCR and COLIEE 2025 benchmarks (Nigam et al., 2025). Compared to these, ViDRILL introduces two distinct contributions: (i) a dual-level chunking strategy optimized independently for retrieval and re-ranking, and (ii) a two-stage hard negative mining mechanism. These innovations enhance semantic robustness and precision in handling long and structured Vietnamese legal texts, thus distinguishing ViDRILL from existing multi-stage frameworks.

Dataset	Total (rows)	Max (words)	Min (words)	Avg (words)
Train (Articles)	59,636	55,097	5	303.28
Train (Questions)	2,190	45	6	19.71
Public Test (Questions)	312	42	11	20.00
Private Test (Questions)	627	57	10	24.40

Table 1: Key statistics of the VLSP 2025 DRiLL dataset. Total indicates the number of rows, while Max/Min/Avg are measured in words.

3 Dataset

In this work, we utilize the dataset from the VLSP 2025 DRiLL shared task (Vuong et al., 2025), which focuses on legal document retrieval in the Vietnamese statutory domain. The dataset consists of a comprehensive legal corpus and an annotated set of questions, building upon prior high-quality Vietnamese legal benchmarks such as VLQA (Nguyen et al., 2025). Each statutory article is identified by a unique `aid` and its full content, while questions are collected from public consultation platforms and annotated by legal professionals with the relevant `aids`. Relevance is defined such that an article is considered "relevant" if the question can be affirmatively answered or entailed from its content. On average, most questions are linked to a single relevant article, though some are associated with multiple. Key dataset statistics, including the maximum length in terms of number of words, are summarized in Table 1.

4 Methodology

We first provide an overview of the proposed framework as illustrated in Figure 1. The system is designed as a multi-stage retrieval pipeline that in-

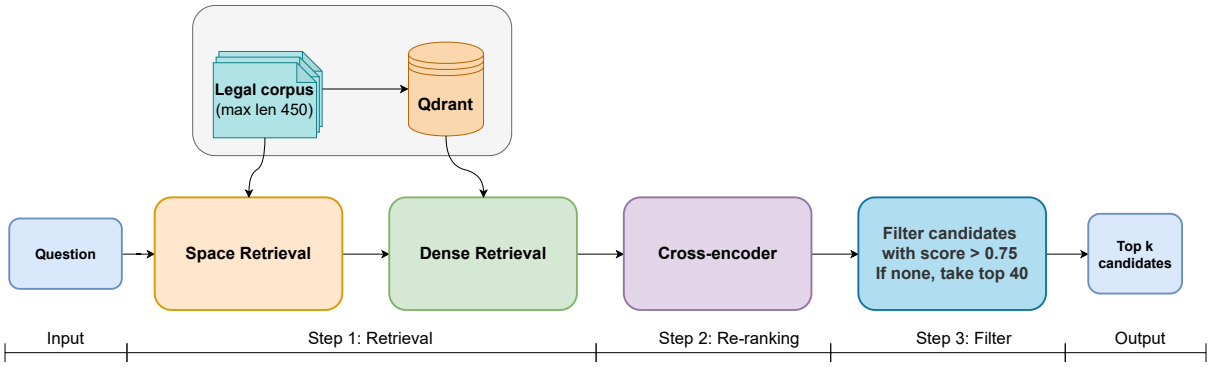

Figure 1: Overall pipeline of our proposed method.

tegrates keyword-based, dense semantic, and cross-encoder models. The pipeline consists of three main stages: retrieval, re-ranking, and filtering.

4.1 Preprocessing

4.1.1 Corpus Preparation

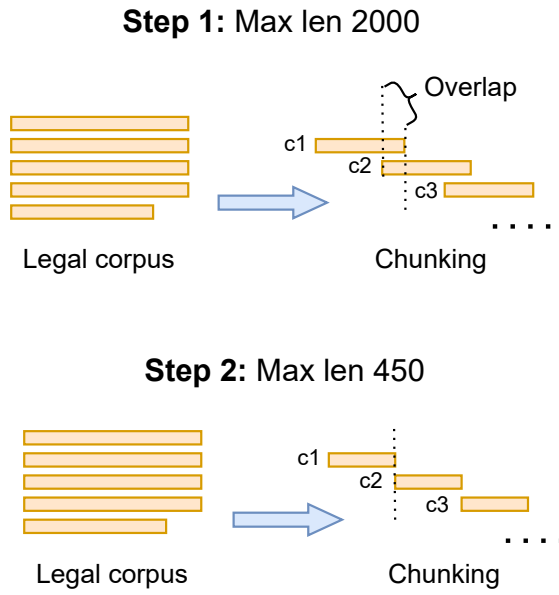

Figure 2: Preprocessing pipeline for corpus segmentation in ViDRILL

The preprocessing stage plays a crucial role in preparing the legal corpus for both retrieval and re-ranking tasks. To ensure high-quality data, we design a two-step chunking strategy with different granularity levels for distinct purposes:

- **Step 1 - Long Chunking for Re-ranking:** Legal documents are segmented into chunks of up to 2000 characters. To maintain semantic continuity across sections, overlapping segments are introduced by preserving content split at line breaks (`\n`). These long chunks are mainly used for training re-ranking models, as they provide sufficient context for the cross-encoder to assess relevance effectively.

- **Step 2 - Short Chunking for Retrieval:** For retrieval models, the corpus is further segmented into smaller chunks with a maximum length of 450 characters. No overlap is applied in this step. Each legal clause is initially treated as an independent unit. In cases where a clause exceeds the character limit, it is further split into smaller coherent segments, following the clause-based segmentation strategy proposed in (Tran et al., 2025). This step ensures compact, self-contained units suitable for dense retrieval models.

This dual-level chunking strategy allows re-ranking models to leverage rich contextual information while ensuring that retrieval models operate on concise and focused textual units. Figure 2 illustrates the preprocessing pipeline applied in ViDRILL.

4.1.2 Training Data Preparation

To prepare high-quality datasets for both retrieval and re-ranking training, we design strategies for constructing positive and negative samples, as well as corpus splits tailored to each model. The details are as follows:

- **Positive Samples:** Each legal article (`aid`) can contain multiple chunks after preprocessing. Instead of treating all chunks as equally positive, we select the most representative one. Specifically, we use the combined

similarity outputs of three pretrained models (BGE-M3 (Chen et al., 2024), E5-Instruct (Wang et al., 2024), and GTE (Zhang et al., 2024)) to rank all candidate chunks for a given aid. The highest-scoring chunk is chosen as the *representative positive*, ensuring that the positive passage used in training is both semantically aligned with the query and contextually faithful to the source document. This selection process avoids noisy positives and improves retrieval precision.

- **Negative Samples for Re-ranking:** To train the re-ranker effectively, we construct *hard negatives* using a two-stage process. First, three pretrained dense models (BGE-M3 (Chen et al., 2024), E5-Instruct (Wang et al., 2024), and GTE (Zhang et al., 2024)) are used to retrieve the top-60 candidate passages for each training query. Passages ranked 10 to 40 are then selected as *preliminary negatives*. These preliminary negatives are used to train a BGE-M3 (Chen et al., 2024) model for retrieval on chunks up to 2000 characters (see Section 4.1.1). After training, the BGE-M3 model is applied to the corpus to retrieve a new top-60 candidate set, which is directly used as the *true hard negatives* for re-ranker training. This two-stage approach ensures that the negatives are both semantically challenging and contextually relevant, enabling the re-ranker to learn fine-grained semantic distinctions.

- **Retrieval Training Data:** For dense retrieval models (E5-Instruct (Wang et al., 2024) and GTE (Zhang et al., 2024)), we use short chunks with length capped at 450 characters. These chunks are compact and self-contained, which makes them well-suited for embedding into fixed-size dense vectors. Instead of manually constructing negatives, we rely on *contrastive learning with in-batch negatives*, where each query's positive passage is contrasted against all other passages in the same batch. This approach scales efficiently and leverages the natural diversity of the legal corpus, making explicit negative sampling unnecessary.

4.2 Data and Question Space

We construct a Vietnamese legal corpus where each law and article may contain multiple clauses. The corpus is indexed in two formats: (i) a sparse index using BM25, and (ii) a dense index stored in Qdrant, a vector database. Sparse indexing provides strong lexical recall, while dense indexing supports semantic similarity matching between queries and passages. The question space follows the ViDRILL task setup, where queries are real-world legal questions that often require nuanced semantic understanding.

4.3 Step 1: Retrieval

The retrieval stage combines both sparse retrieval (BM25) and dense retrieval (E5-Instruct (Wang et al., 2024), GTE (Zhang et al., 2024), and BGE-M3). First, BM25 retrieves the top-200 passages, ensuring high recall. Next, dense retrievers encode queries and short chunks (length $\leq$ 450 characters) into embeddings. Qdrant is then used for fast similarity search based on cosine similarity. Each dense model retrieves its top-30 candidates, and their union is taken with duplicates removed to form a unified candidate pool of semantically relevant passages.

4.4 Step 2: Re-ranking

For re-ranking, we adopt the BGE-reranker-v2-m3 (Chen et al., 2024), a multilingual cross-encoder optimized for passage ranking. The re-ranker takes a query and a candidate passage (long chunk up to 2000 characters) as input and outputs a scalar relevance score. Training is performed using positive passages paired with hard negatives (as constructed in the previous section), with a softmax cross-entropy loss computed over each candidate group. This stage leverages full cross-attention between queries and passages, significantly improving ranking precision by capturing fine-grained semantic relationships.

4.5 Step 3: Filtering and Final Selection

Finally, we apply a filtering mechanism to produce the system output. From the re-ranked list, candidates with scores ≥ 0.75 are retained. If there are no candidates satisfying this threshold, the top-40 passages are selected by default. This strategy ensures a balance between precision (via score thresholding) and recall (via fallback selection), ensuring robust performance across diverse queries.

5 Experimental Setup

5.1 Implementation Details

We design a multi-stage pipeline consisting of a retrieval model and a re-ranking model. Table 2 summarizes the implementation details of each component.

Configuration	Retrieval	Re-ranking
Training objective	MultiNegLoss	Cross-Entropy Loss
Max seq. length	512	P=512 / Q=128
Batch / Group size	16	10
Learning rate	2×10^{-5}	6×10^{-5}
Epochs	30	10

Table 2: Implementation details for retrieval and re-ranking modules.

5.2 Evaluation Metrics

System performance on the retrieval task is evaluated using **Precision**, **Recall**, and the **macro $\mathbf{F}_2$ score**. We adopt the macro-average setting, i.e., the evaluation measure is first computed individually for each query and then averaged across all queries.

Precision.

$$\text{Precision} = \frac{1}{|Q|}\sum_{q\in Q}\frac{|R_q \cap G_q|}{|R_q|}$$

where R_q is the set of retrieved documents for query q and G_q is the set of relevant documents. Precision measures the proportion of retrieved documents that are relevant (those that are also in G_q).

Recall.

$$\text{Recall} = \frac{1}{|Q|}\sum_{q\in Q}\frac{|R_q \cap G_q|}{|G_q|}$$

Recall measures the proportion of relevant documents that are successfully retrieved.

Macro $\mathbf{F}_2$ measure.

$$F_2 = \frac{5 \times \text{Precision} \times \text{Recall}}{4 \times \text{Precision} + \text{Recall}}$$

The F_2 score gives recall $4\times$ more weight than precision, which aligns with the legal retrieval scenario where retrieving all relevant precedents is more important than achieving high precision.

6 Results and Analysis

6.1 Experimental Results on Public Test

Table 3 provides several noteworthy observations. First, dense retrieval models with re-ranking consistently outperform the single baselines. For example, the E5-Instruct model, when combined with re-ranking, achieves an improvement of approximately 0.13–0.15 in F_2-Macro compared to the lexical baseline BM25 and the dense model GTE without re-ranking. This demonstrates the effectiveness of semantic representations in capturing legal relevance beyond surface lexical overlap.

Second, in Step 1 of the Single+Re-rank configuration, increasing the retrieval pool size (top-k) generally improves recall while slightly reducing precision. Among the examined settings, $k = 30$ achieves the most balanced trade-off, yielding an F_2-Macro of 0.6641. This finding indicates that enlarging the candidate pool improves coverage of relevant legal documents, but excessively large values of k introduce noise that harms precision. Consequently, we select $k = 30$ as the default retrieval depth for subsequent experiments.

Third, in Step 2, where the retrieval depth is fixed at $k = 30$, varying the fallback range reveals its impact on robustness. Allowing the system to retain up to 40 candidates (top-40) provides the most favorable balance, with an F_2-Macro of 0.7052. This suggests that fallback strategies play a crucial role in avoiding the loss of relevant documents when re-ranker confidence is low, thereby improving recall stability.

Finally, combining sparse and dense retrieval in the Double+Re-rank setting leads to further improvements. By integrating BM25 with E5-Instruct and GTE, the system reaches an F_2-Macro of 0.7135 and recall up to 0.8080. These results highlight that hybrid retrieval is particularly beneficial in the legal domain, where queries often require both lexical precision and semantic understanding. Overall, the analysis suggests that carefully balancing retrieval depth, fallback strategies, and hybridization is essential for optimizing performance in complex legal search scenarios.

6.2 Private Test Leaderboard

Table 4 presents the top-10 teams on the private test set. Our system (ViDRILL) ranks 5th, distinguished by a strong recall of 0.7605, among the top 2 across all teams. This indicates that our system is effective at retrieving a broad set of relevant legal

Retrieval			Re-ranking	Filter		Result		
Sparse	**Dense**	**Top-k**		**Threshold**	**Fallback**	**F2-Macro**	**Precision**	**Recall**
(a) Single: baseline without re-ranking								
BM25	–	50	✗	–	top-02	0.3177	0.2115	0.3632
–	GTE	50	✗	–	top-02	0.5031	0.3397	0.5718
–	E5-Instruct	50	✗	–	top-02	0.5338	0.3622	0.6055
(b) Single + Re-rank (Step 1: analyze top-k with fixed fallback = top-01)								
–	E5-Instruct	10	✓	0.75	top-01	0.6583	**0.5877**	0.6787
–	E5-Instruct	30	✓	0.75	top-01	0.6641	0.5631	0.6952
–	E5-Instruct	50	✓	0.75	top-01	0.6626	0.5517	0.6976
(c) Single + Re-rank (Step 2: fix top-k = 30, vary fallback strategy)								
–	E5-Instruct	30	✓	0.75	top-10	0.6947	0.5020	0.7684
–	E5-Instruct	30	✓	0.75	top-20	0.7018	0.4938	0.7845
–	E5-Instruct	30	✓	0.75	top-40	0.7052	0.4920	0.7909
(d) Double + Re-rank (hybrid retrieval BM25 + dense)								
BM25	E5-Instruct+GTE	30	✓	0.75	top-20	0.7055	0.4886	0.7935
BM25	E5-Instruct+GTE	30	✓	0.75	top-40	**0.7135**	0.4861	**0.8080**
BM25	E5-Instruct+GTE	30	✓	0.75	top-50	0.7132	0.4854	**0.8080**

Table 3: Experimental results on the public test set. (a) *Single* shows baseline retrieval without re-ranking. (b) *Single + Re-rank Step 1* varies the top-k retrieval (with fixed fallback = top-01), showing that top-30 achieves the best trade-off. (c) *Single + Re-rank Step 2* fixes top-k=30 and varies fallback size, where top-40 is the most effective. (d) *Double + Re-rank* combines BM25 with dense retrievers, further improving recall.

Team	F2-Macro	Precision	Recall
edmmm	**0.7261**	**0.6773**	0.7394
unknow_123	0.6966	0.6222	0.7181
ducanger	0.6955	0.5097	**0.7653**
dinhanhx	0.6710	0.5509	0.7097
ViDRILL (ours)	0.6521	0.4153	0.7605
truong13012004	0.6495	0.4714	0.7172
AImba	0.6425	0.4086	0.7498
ngjabach	0.6280	0.4329	0.7077
Engineers	0.5864	0.3147	0.7478
villageai	0.5587	0.3799	0.6332

Table 4: Top-10 results on the private test set of VLSP 2025 DRiLL.

documents, which is crucial in legal applications where missing information must be minimized. However, precision remains lower (0.4153), suggesting that future improvements should focus on re-ranking and filtering strategies to reduce irrelevant retrievals while maintaining high recall.

7 Discussion

Using different retrieval methods together with re-ranking helps balance precision and recall. Recall-focused filtering, including thresholding and fallback strategies, ensures important legal documents are not missed. Choosing an appropriate threshold is crucial: too low may include irrelevant documents, while too high may miss relevant ones. Adding too many dense retrievers can create noise and reduce accuracy. From this work, we learned that balancing coverage and precision, setting proper thresholds, and selecting models based on their strengths are key for effective legal retrieval.

8 Conclusion

ViDRILL addresses the challenging task of Vietnamese legal document retrieval, characterized by complex, lengthy texts and nuanced queries. Our key contributions include a dual-level chunking strategy for optimized preprocessing and a two-stage hard negative mining mechanism to enhance semantic robustness. By integrating BM25, dense multilingual encoders (E5-Instruct, GTE), and BGE-reranker-v2-m3, our multi-stage framework achieves top-5 performance in the VLSP 2025 DRiLL shared task. This work advances Vietnamese legal NLP by demonstrating the efficacy of hybrid retrieval pipelines, offering a scalable and effective solution for real-world legal search applications.

Future enhancements to ViDRILL will focus on integrating large language models (LLMs) for advanced re-ranking, leveraging their contextual understanding to improve ranking precision. We aim to refine retrieval by optimizing model combinations and exploring domain-specific fine-tuning to

better capture legal nuances. Additionally, incorporating query reformulation and dynamic thresholding strategies will enhance adaptability to diverse query types, further improving both precision and recall for practical legal retrieval systems.

References

Sophia Althammer, Arian Askari, Suzan Verberne, and Allan Hanbury. 2021. Dossier@coliee 2021: Leveraging dense retrieval and summarization-based reranking for case law retrieval. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAIL) 2021*.

Arian Askari, Georgios Peikos, Gabriella Pasi, and Suzan Verberne. 2022. Leibi@coliee 2022: Aggregating tuned lexical models with a cluster-driven bert-based model for case law retrieval. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAIL) 2022*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Yi Feng, Chuanyi Li, and Vincent Ng. 2024. Legal case retrieval: A survey of the state of the art. In *Proceedings of the 62nd Annual Meeting of the ACL (Long Papers)*, pages 6472–6485. Association for Computational Linguistics.

Haitao Li, Qingyao Ai, Xinyan Han, Jia Chen, and et al. 2024. Delta: Pre-train a discriminative encoder for legal case retrieval via structural word alignment. *arXiv preprint*. Structural word alignment for legal retrieval.

Yixiao Ma, Yueyue Wu, Weihang Su, Qingyao Ai, and Yiqun Liu. 2023. Caseencoder: A knowledge-enhanced pre-trained model for legal case encoding. In *Proceedings of [specific conference]*. Biased Circle Loss for case relevance encoding.

Chau Nguyen, Nguyen-Khang Le, Dieu-Hien Nguyen, Phuong Nguyen, and Le-Minh Nguyen. 2022. A legal information retrieval system for statute law. In *Proceedings of ACIIDS 2022, CCIS*, volume 1716, pages 370–382. Springer.

Hai-Long Nguyen, Tan-Minh Nguyen, Duc-Minh Nguyen, Thi-Hai-Yen Vuong, Ha-Thanh Nguyen, and Xuan-Hieu Phan. 2024. Exploiting llms' reasoning capability to infer implicit concepts in legal information retrieval. *arXiv preprint arXiv:2410.12154*.

Tan-Minh Nguyen, Hoang-Trung Nguyen, Trong-Khoi Dao, Xuan-Hieu Phan, Ha-Thanh Nguyen, and Thi-Hai-Yen Vuong. 2025. Vlqa: The first comprehensive, large, and high-quality vietnamese dataset for legal question answering.

Shubham Kumar Nigam, Tanmay Dubey, Noel Shallum, and Arnab Bhattacharya. 2025. Segment first, retrieve better: Realistic legal search via rhetorical role-based queries. In *COLIEE 2025*. IL-PCR and COLIEE 2025.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Son Pham Tien, Hieu Nguyen Doan, An Nguyen Dai, and Sang Dinh Viet. 2024. Improving vietnamese legal document retrieval using synthetic data. In *Proceedings of the 10th International Workshop on Vietnamese Language and Speech Processing (VLSP 2024)*. Synthetic queries generation for retrieval fine-tuning.

Dien X. Tran, Nam V. Nguyen, Thanh T. Tran, Anh T. Hoang, Tai V. Duong, Di T. Le, and Phuc-Lu Le. 2025. Semviqa: A semantic question answering system for vietnamese information fact-checking.

Thi-Hai-Yen Vuong, Tan-Minh Nguyen, Hoang-Trung Nguyen, Trong-Khoi Dao, Ha-Thanh Nguyen, and Hoang-Quynh Le. 2025. Overview of the vlsp 2025 challenge on drill: Deep retrieval in the expansive legal landscape. In *Proceedings of the 11th International Workshop on Vietnamese Language and Speech Processing*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.