Data Augmentation and Hierarchical Chunking for Deep Retrieval in the Expansive Legal Landscape

Le Ba Hoai
AI R&D, FPT IS
lebahoaidongson@gmail.com

Dinh Van Hung
AI R&D, FPT IS
hungdv59@fpt.com

Tran Bao Hieu
AI R&D, FPT IS
hieutb2@fpt.com

Vu Dinh Anh
AI R&D, FPT IS
anhvd27@fpt.com

Pham Quang Nhat Minh
AI R&D, FPT IS
minhpqn@fpt.com

Abstract

Legal retrieval in the DRiLL 2025 competition requires addressing the inherent complexity and hierarchical nature of legal documents. Traditional preprocessing often produces overly long or incoherent chunks, leading to contextual loss during retrieval. To overcome these limitations, we propose a structured data transformation pipeline that combines LLM-based data augmentation and hierarchical chunking. First, large language models are employed to augment and enrich the raw legal texts, providing more diverse and semantically consistent representations. Next, hierarchical chunking decomposes large sections into smaller, well-formed units that preserve logical flow while improving granularity. This two-step process ensures that the data is both semantically richer and structurally coherent, forming a stronger foundation for downstream retrieval models. Our main contribution lies in demonstrating that structured preprocessing—through augmentation and hierarchical segmentation—significantly enhances retrieval quality in expansive legal corpora. The approach provides a scalable and adaptable framework that can be integrated into various retrieval pipelines for complex domains such as law.

Keywords— legal information retrieval, semantic search, hierarchical chunking, embedding, reranking, Vietnamese NLP

1 Introduction

The expeditious evolution of artificial intelligence, particularly generative models in NLP such as those exemplified by advanced LLMs, has amplified the imperative for sophisticated tools in legal text processing (Team, 2024). Legal NLP encompasses a spectrum of tasks including document classification, entity recognition, and information retrieval, each tailored to navigate the intricacies of legal language characterized by domain-specific jargon, long dependencies, and contextual nuances (Malik et al., 2024). Although substantial advancements have characterized Legal NLP in languages including English, Japanese, and Chinese, foundational inquiries into Vietnamese legal text processing remain nascent, hampered by limited annotated datasets and linguistic complexities such as tonal variations and compound words (Lefterov et al., 2023). In this context, the VLSP 2025 DRiLL shared task introduces a pioneering initiative to propel Vietnamese Legal NLP forward by providing a benchmark dataset for legal document retrieval (Nguyen et al., 2025).

IR represents a pivotal NLP endeavor, centered on pinpointing query-relevant information with high precision and recall (Zhao et al., 2022). In legal contexts, this manifests as retrieving articles germane to queries, formalized as subset identification predicated on semantic relevance. Traditional IR methods, such as term frequency-inverse document frequency (TF-IDF) or BM25, often falter in capturing semantic depth, particularly in multilingual settings like Vietnamese where synonyms and polysemy abound (Zhao et al., 2022). Dense retrieval paradigms, leveraging embeddings from transformer-based models, have emerged as alternatives, yet they grapple with scalability issues in voluminous legal corpora (Zhao et al., 2022).

The dataset encompasses over 2000 Vietnamese legal documents segmented into articles, albeit frequently bereft of titular metadata, engendering contextual deficits that exacerbate retrieval inaccuracies. Contemporary retrieval paradigms leveraging dense embeddings and semantic search encounter impediments with protracted texts and noisy datasets under resource constraints, including token length limitations and computational overhead (Zha et al., 2023). To surmount these, we espouse a data-centric hybrid paradigm, accentuating preprocessing for metadata reconstruction, chunking for embedding tractability, and multistage retrieval with noise attenuation. This paradigm draws

conceptual parallels from hybrid methodologies in analogous NLP tasks, such as multi-document summarization, adapted herein to retrieval exigencies (Malik et al., 2024). By centering on data quality enhancement rather than model parameter tuning, our approach aligns with emerging trends in data-centric AI, which emphasize iterative data refinement to boost downstream performance (Zha et al., 2023).

2 Related Work

Legal IR has been extensively studied in monolingual English contexts, with benchmarks like COLIEE highlighting the efficacy of dense retrievers and rerankers (Rabelo et al., 2024). For Vietnamese, recent efforts include multi-stage retrieval frameworks that integrate lexical and semantic methods to handle legal texts' complexity (Lefterov et al., 2023). Data-centric approaches in NLP have gained traction, shifting focus from model architecture to data curation, as evidenced in surveys advocating for systematic data engineering to improve task-specific performance (Zha et al., 2023). Hierarchical chunking techniques have been proposed to manage long documents in retrieval-augmented generation (RAG) systems, enabling granular yet context-aware segmentation (Li et al., 2025a). Multilingual embeddings, such as those from BGE-M3, have demonstrated versatility in cross-lingual IR, including legal domains (Chen et al., 2024). Reranking models enhance initial retrieval by refining candidate rankings, with applications in legal QA systems (Clavié et al., 2024). Clustering for noise reduction in multi-vector pipelines has been explored to filter irrelevant representations, improving overall retrieval robustness (Podolny et al., 2025). Vector databases like Milvus support scalable NLP applications by enabling efficient similarity searches (Wang et al., 2021). Our work builds on these by integrating them into a cohesive data-centric pipeline for Vietnamese legal IR.

3 Methodology

Our paradigm is inherently data-centric, privileging data integrity and manipulation over model fine-tuning, thereby reducing dependency on extensive computational resources (Zha et al., 2023). It encompasses phases of data reconstruction, hierarchical chunking, embedding and archival, semantic retrieval with reranking, and post-processing via clustering and LLM adjudication. The pipeline schematic is analogous to hybrid architectures in extant literature, where initial extraction precedes abstractive refinement.

3.1 Data Preprocessing

Our dataset consists of over 2,000 legal documents that were split into articles. However, many articles lack titles, which reduces the clarity and semantic context of the text. To address this, we reconstruct the data by generating missing titles and producing summaries with a large language model (LLM). Specifically, we use Qwen2.5-7B-Instruct, an instruction-tuned model with strong multilingual ability (including Vietnamese) and robust reasoning performance (Team, 2024). This process produces concise summaries and inferred titles, enriching the dataset with additional metadata and improving its overall quality (Lefterov et al., 2023).

3.2 Hierarchical Chunking

After reconstruction, many articles remain very long (often more than 10,000 tokens), which exceeds the input limits of most open-source embedding models (typically around 4,000 tokens). To make the data usable, we apply hierarchical chunking (Li et al., 2025b). First, the generated article title is treated as the root node. We then split the article into smaller units (e.g., clauses or bullet points), building a hierarchical tree structure. The leaf nodes of this tree are used as chunks. Each chunk is enriched with metadata describing its path from the root to the leaf, preserving structural context. If a leaf node still exceeds 4,000 tokens, we further split it into overlapping segments (chunk size = 2048 tokens, overlap = 1024). This ensures that all chunks are small enough for embedding while keeping both detail and context.

3.3 Embedding and Storage

We embed the chunks using BGE-M3, a multilingual embedding model designed for dense retrieval across 100+ languages, including Vietnamese (Chen et al., 2024). Each chunk is mapped to a 1024-dimensional vector that captures its semantic meaning. The embeddings are stored in Milvus, a vector database optimized for large-scale similarity search, which supports efficient approximate nearest neighbor (ANN) search with HNSW indexing (Wang et al., 2021). Metadata such as hierarchy paths are stored alongside the embeddings

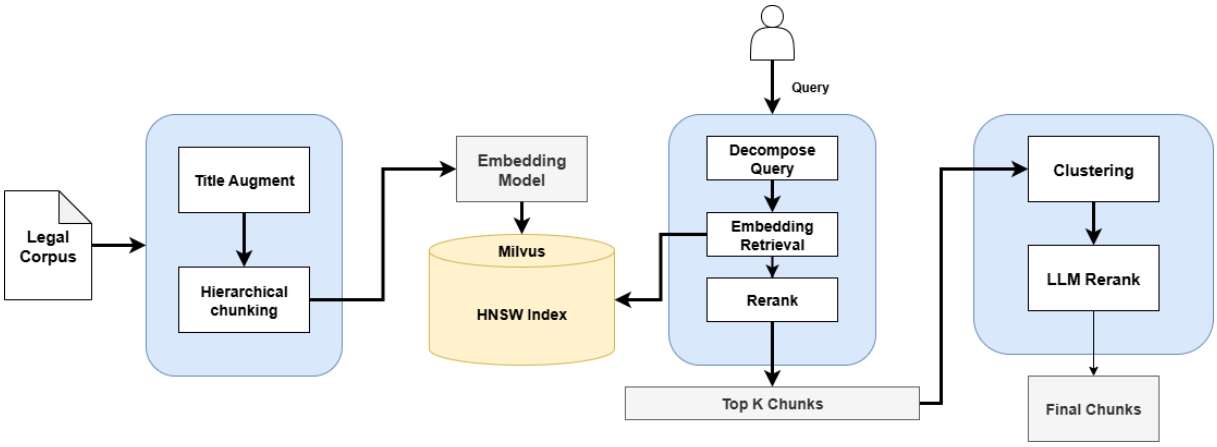

Figure 1: End-to-end system architecture

to improve retrieval accuracy. In our setup, Milvus handles large volumes of embeddings with query latencies below 50ms.

3.4 Retrieval and Reranking

When a query is issued, we first use an LLM to decompose the query into steps. For each step, candidate chunks are retrieved from Milvus using cosine similarity:

- $\text{sim}(q, d) = \frac{q \cdot d}{|q||d|}$

where q and d are the query and document embeddings, respectively (Zhao et al., 2022). We initially retrieve the top-50 candidates, which are then reranked using BGE-Reranker-v2-M3, a compact cross-encoder that scores chunk–query pairs (Clavié et al., 2024). Scores are normalized to the [0,1] range, and only chunks with a score above 0.7 are kept. This reranking step improves precision by reducing false positives.

3.5 Post-Processing

Reranking can still introduce noise or redundant results. To mitigate this, we: 1. Deduplicate by article ID, keeping only the entry with the highest score. 2. Cluster the remaining candidates by their rerank scores using K-means, with at most 7 clusters. For single-item cases, we assign a default cluster. 3. Compute cluster statistics (centroid, variance, score range) and sort clusters by centroid score. 4. Select the top-5 clusters for final evaluation.

The final evaluation is done with Qwen2.5-72B using chain-of-thought prompting to ensure semantic relevance. This step acts as a zero-shot entailment check, filtering the most relevant chunks for the user query (Podolny et al., 2025).

3.6 Implementation

We run embeddings and reranking on a single NVIDIA A30 GPU (12GB VRAM). LLM inference (Qwen2.5 variants) is handled through a third-party API and run with 4-bit quantization for efficiency (Team, 2024). Processing the full dataset (embedding and storage) takes about 2 hours, with Milvus configured using HNSW parameters $M = 128$ and $efConstruction = 200$. In practice, query retrieval (including reranking and post-processing) is completed within 100ms.

3.7 Experimental Results

3.7.1 Deep Retrieval in the Legal Landscape Challenge Dataset

We use the dataset released as part of the **VLSP 2025 DRiLL** (Deep Retrieval in the Legal Landscape) shared task. The dataset is designed to evaluate systems on retrieving relevant legal documents given a natural language query.

Corpus. The legal corpus consists of Vietnamese law texts, including statutes, decrees, and other regulations. Each article in the corpus is assigned a unique identifier and serves as a candidate unit for retrieval. Legal articles vary substantially in length, ranging from short provisions to long multi-clause documents with thousands of tokens.

Splits. The organizers released a *training set* (pairs of queries and relevant legal articles) on July 1, 2025, and a *public test set* on July 15, 2025. The public test set contains 10,000 unlabeled queries used for leaderboard ranking. A *private test set* with hidden labels was later released for the final evaluation phase on August 8, 2025.

Task definition. The retrieval task is defined as follows: given a legal query q, the system must return a subset of articles $A' \subset A$ such that the query is entailed by or answered within each article in A'. Each query may map to one or multiple relevant articles. Query lengths vary from approximately 4 to 73 tokens, with an average length of around 20 tokens, posing challenges for representation learning and retrieval effectiveness.

Constraints. Participants are not allowed to use external labeled datasets. However, they may employ publicly released pretrained models or corpora available prior to January 1, 2025. Closed-source or proprietary LLMs (e.g., GPT-4, Gemini) are disallowed to ensure fairness and reproducibility.

Evaluation. Performance on both the public and private test sets is assessed using retrieval metrics such as *Precision*, *Recall*, and F_2-*Macro* (detail defined below). Leaderboard ranking is determined by these metrics, with F_2-Macro serving as the primary measure, balancing retrieval precision and coverage.

- Precision = average of (correctly retrieved articles per query) / (retrieved articles per query)
- Recall = average of (correctly retrieved articles per query) / (relevant articles per query)
- F2 = (5 $\times$ Precision $\times$ Recall) / (4 $\times$ Precision + Recall)

3.7.2 Impact of Metadata Augmentation and Hierarchical Chunking

Metric	aug and chunk	w/o aug and chunk
Precision	**0.7981**	0.4341
Recall	**0.7981**	0.4992
F2-Macro	**0.7426**	0.4847

Table 1: Performance metrics on the public leaderboard (with and without augmentation and hierarchical chunking).

Table 1 presents the ablation study on the effect of metadata augmentation and hierarchical chunking. The results show a clear improvement across all evaluation metrics. With augmentation and chunking, both Precision and Recall reach 0.7981, representing substantial gains over the baseline without augmentation (0.4341 in Precision and 0.4992 in Recall). This consistent improvement leads to a remarkable increase in the overall $F2 - Macro$ score (0.7426 vs. 0.4847). Such results demonstrate that enriching the dataset with metadata and applying hierarchical chunking not only enhance the semantic representation of legal texts but also significantly boost retrieval effectiveness.

3.7.3 Our Result at DRiLL Challenge

Table 2 shows the leaderboard of the Drill Challenge. Our results achieved the second rank with $F2-Macro = 0.6966$, which highlights the effectiveness of our proposed method. In terms of Precision, our method reached 0.6222, exceeding the third-ranked team by a large margin of 0.13. This indicates that our approach can significantly reduce false positives compared to competing methods, while maintaining a competitive Recall of 0.7181.

4 Limitations

A primary limitation of our approach stems from the reliance on LLMs for generating titles and summaries during the data reconstruction phase. Specifically, the titles and summary content are autonomously produced by the LLM (Qwen2.5-7B-Instruct) based on the original article content, rendering the accuracy of these augmentations heavily contingent upon the model's generative fidelity. In legal domains, where precision is paramount, LLMs are susceptible to hallucinations—generating plausible but factually erroneous information—or generalization biases, such as omitting critical qualifiers that limit the scope of legal provisions, potentially leading to overbroad interpretations. This dependency may introduce subtle inaccuracies in the reconstructed metadata, which could propagate through the embedding and retrieval stages, affecting overall system reliability. Furthermore, while our method mitigates some issues through hierarchical chunking and reranking, it does not fully address potential biases inherent in the LLM's training data, particularly for underrepresented languages like Vietnamese in legal contexts. Future iterations could incorporate human-in-the-loop validation or ensemble LLMs to enhance robustness.

5 Conclusion

This exposition articulates a data-centric resolution for the VLSP 2025 Vietnamese Legal Information Retrieval task. Through meticulous data reconstruction, hierarchical chunking, and multi-

#	Participant	ID	F2-Macro	Precision	Recall
1	edmmm	352904	**0.7261**	0.6773	0.7394
2	**our method**	352595	0.6966	0.6222	0.7181
3	ducanger	352643	0.6955	0.5097	0.7653
4	dinhanhx	352240	0.6710	0.5509	0.7097
5	fasterunited	352352	0.6521	0.4153	0.7605
6	truong13012004	352102	0.6495	0.4714	0.7172
7	Almba	351474	0.6425	0.4086	0.7498
8	ngjbach	352906	0.6280	0.4329	0.7077
9	Engineers	351851	0.5864	0.3147	0.7478
10	villageai	351120	0.5587	0.3799	0.6332
11	quintu	352474	0.5578	0.2769	0.7472

Table 2: Leaderboard results at Drill Challenge.

stage retrieval encompassing embedding, reranking, and clustering, competitive efficacy is realized absent fine-tuning. The approach not only addresses dataset deficiencies but also scales to real-world legal corpora, offering a blueprint for multilingual IR. Prospective endeavors may encompass domain-specific embedding fine-tuning, integration of advanced clustering algorithms like density-based methods, or extension to multimodal legal documents incorporating tables and figures (Zha et al., 2023).

References

Jianlv Chen, Shitao Xiao, Peitian Zhang, and 1 others. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Nicolas Clavié, Amin Dadgar, Guillaume Chamelot, Paul Fiquet, Aymen Shabou, and Hervé Déjean. 2024. Enhancing qa text retrieval with ranking models: Benchmarking, fine-tuning and deploying rerankers for rag. *Preprint*, arXiv:2409.07691.

Daniel Lefterov, George Tsakalidis, Ian Soboroff, and Christina Lioma. 2023. Natural language processing in the legal domain. *Preprint*, arXiv:2302.12039.

C. Li and 1 others. 2025a. Enhancing retrieval augmented generation with hierarchical text segmentation and clustering. *Preprint*, arXiv:2507.09935.

Y. Li and 1 others. 2025b. Hierarchical document refinement for long-context retrieval-augmented generation. *Preprint*, arXiv:2505.10413.

M. S. Malik, Sridhar Gopikrishnan, and Praveen Paritosh. 2024. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *Preprint*, arXiv:2410.21306.

Tan-Minh Nguyen, Hoang-Trung Nguyen, and 1 others. 2025. Vlqa: The first comprehensive, large, and high-quality vietnamese dataset for legal question answering. *Preprint*, arXiv:2507.19995.

A. S. Podolny and 1 others. 2025. Crisp: Clustering multi-vector representations for denoising and pruning. *Preprint*, arXiv:2505.11471.

Juliano Rabelo, Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Ken Satoh, and Masaharu Yoshioka. 2024. Overview of benchmark datasets and methods for the legal information extraction and entailment competition (coliee 2024). In *New Frontiers in Artificial Intelligence - JURISIN 2024 Workshops*, pages 104–118.

Qwen Team. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Jianguo Wang, Tianyi Li, Xiaomeng Yi, Jing Xu, M. Tamer "Ozsu, Zikun Deng, Rubin Mao, Weijie Zhao, Xiaohu Cai, Lu Qiu, Fei Xue, Jingjing Wu, Zhuoer Feng, Jing Wang, Yuntao Li, and Wei Lin. 2021. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2614–2627.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey. *Preprint*, arXiv:2303.10158.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2022. Dense text retrieval based on pretrained language models: A survey. *Preprint*, arXiv:2211.14876.