EDM Team's Approach in DRiLL@VLSP 2025: A Multi-Stage End-to-End Architecture for Deep Retrieval in the expansive Legal Landscape

Tien-Duy Pham, Huu-Dong Nguyen, The-Ngoc Phung
Electronics Communication Technology Investment Development Corporation, Hanoi, Vietnam
{duypt, dongnh, ngocpt}@elcom.com.vn

Abstract

The rapid development of large language models has advanced legal natural language processing in high-resource languages such as English and Chinese, whereas research for Vietnamese remains limited. In this work, a system is presented for the Vietnamese legal document retrieval task in the VLSP 2025 DRiLL Challenge. A multi-stage retrieval pipeline is designed, combining BM25, dense embeddings, neural rerankers, and prompting with large language models. To improve effectiveness, automatically generated article titles and inter-article reference features are incorporated. A learning-to-rank framework is employed to integrate these signals and progressively refine candidate sets from 500 to 100 and finally to 10, thereby preserving recall while enhancing precision. Evaluation on the private test set of the VLSP 2025 DRiLL Challenge shows that the proposed system achieved the highest reported performance, with an F2 score of 0.7261, Precision of 0.6773, and Recall of 0.7394. These findings confirm the effectiveness of hybrid retrieval and feature enrichment for Vietnamese legal texts and provide a foundation for future research in low-resource legal natural language processing.

1 Introduction

Legal question answering has become an important research direction in natural language processing (NLP), where information retrieval plays a crucial role in identifying relevant articles that serve as the foundation for answer extraction. Legal documents are typically long, complex in structure, and heavily interconnected, which makes retrieval a particularly challenging task for low-resource languages such as Vietnamese. While recent progress in large language models has advanced legal NLP in high-resource languages, Vietnamese legal text processing remains underdeveloped.

To address this gap, the VLSP 2025 Challenge on DRiLL: The Challenge of Deep Retrieval in the Expansive Legal Landscape was organized. The task requires determining which legal articles are relevant to a given question, where relevance is defined by whether the article entails a yes-or-no answer. The dataset consists of over two thousand Vietnamese legal documents with nearly sixty thousand articles, accompanied by more than three thousand training and evaluation questions. The challenge enforces strict constraints, permitting only the official dataset and publicly released models before January 2025, ensuring fairness and reproducibility (Vuong et al., 2025).

In this paper, a multi-stage retrieval pipeline is introduced, integrating lexical retrieval, dense embeddings, re-ranking, and prompting with large language models. Additional features such as automatically generated titles and inter-article references are incorporated to capture the structure of legal texts, while a learning-to-rank algorithm refines results across successive stages. The proposed system achieved the best performance in the competition, demonstrating the effectiveness of hybrid retrieval and feature enrichment for Vietnamese legal natural language processing.

While optimized for VLSP 2025 DRiLL, the system was designed for broader applicability. Its architecture and domain-independent features such as embeddings, reranker outputs, LLM-derived scores, and token length enable transfer to other corpora with minimal adjustments, typically limited to Chain-of-Thought (CoT) prompt adaptation. Since no fine-tuning of large pretrained models is required and only lightweight learning-to-rank (LTR) modules are retrained, domain transfer remains efficient and cost-effective.

2 Related Work

Legal information retrieval has long been recognized as challenging due to the length, structural complexity, domain-specific terminology, and

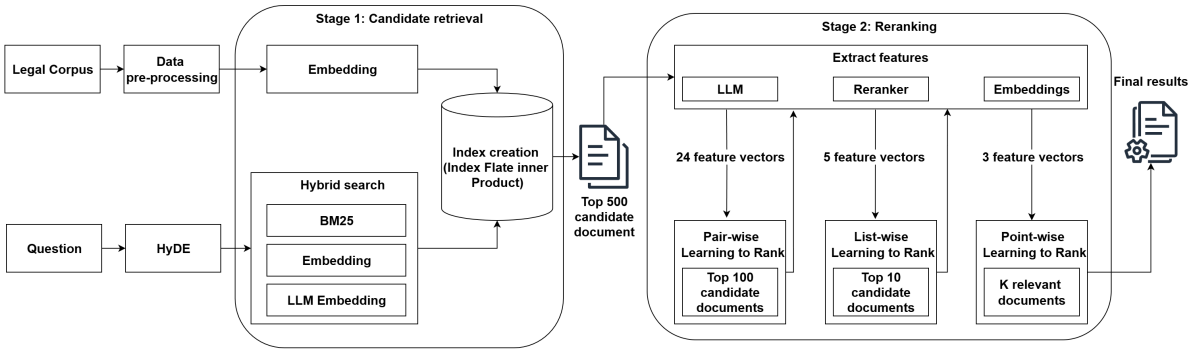

Figure 1: Overall architecture for retrieval legal documents.

dense inter-document dependencies of legal texts (Zhong et al., 2020). Early approaches were predominantly based on lexical matching, with BM25 widely adopted for keyword-oriented retrieval; however, limitations under paraphrasing and semantic variation have been repeatedly documented (Robertson and Zaragoza, 2009).

With the advent of deep learning, dense retrieval was introduced in which queries and documents are encoded into shared vector spaces to enable semantic matching, as exemplified by Dense Passage Retrieval (Karpukhin et al., 2020). Improved legal-domain semantics have been obtained through domain-adapted encoders such as Legal-BERT (Chalkidis et al., 2020). Hybrid methods that combine lexical and dense signals and are refined by cross encoder reranking have shown superior effectiveness, including results on COLIEE (Kano et al., 2021). In parallel, Hypothetical Document Embeddings (HyDE) were proposed: an LLM synthesizes a pseudo document from the query and the embedding of this document is used for retrieval, yielding strong zero shot gains without relevance labels (Gao et al., 2023).

For low resource languages including Vietnamese, slower progress has been observed relative to English and Chinese. New Vietnamese legal datasets have supported monolingual modeling (Nguyen et al., 2023, 2025), and prompting with large language models has been leveraged for zero shot and low shot relevance judgments (Wei et al., 2022). Learning to rank has been employed to integrate heterogeneous indicators, including lexical scores, semantic similarities, and citation based structural features, with LambdaMART widely used in practice (Liu, 2009; Burges, 2010; Wu et al., 2010).

Multi stage pipelines have been emphasized in legal and general IR: fast candidate generation using BM25 is followed by increasingly powerful rerankers, including cross encoders and LLM based scoring (Pradeep et al., 2021). Further gains have been reported when hybrid retrieval is coupled with learning to rank and optimized using listwise objectives that align with ranking metrics (Liu, 2009; Ma et al., 2020). Within the DRiLL challenge context, such multi stage designs have been highlighted for balancing recall and precision over expansive legal corpora (Vuong et al., 2025), and Vietnamese legal applications have explored learning to rank for tasks such as query auto completion and sentence ranking to explain statutory terms (Tran et al., 2022).

3 Method

3.1 Overall Architecture

Figure 1 illustrates the overall multi-stage architecture of our retrieval system, which is designed to balance recall, precision, and efficiency in large-scale Vietnamese legal corpora. The pipeline begins with preprocessing and indexing, where each legal article is represented using both lexical (BM25) and semantic embeddings to support hybrid retrieval. In the candidate retrieval stage, a combination of BM25, dense embeddings, and HyDE-generated queries, together with inter-article references, yields up to 500 highly recalled candidates. These results are then progressively refined through two re-ranking stages: a pairwise LambdaMART model reduces the set to 100 by leveraging heterogeneous lexical, semantic, and structural features, while a listwise LightGBM (Ke et al., 2017) ranker further optimizes global ranking consistency to produce the final top-10 articles. This hierarchical design allows early stages to maximize recall while later stages introduce stronger signals for precision, ensuring robust and accurate retrieval

performance for downstream legal question answering.

3.2 Preprocessing

The Vietnamese Legal Question Answering (VLQA) dataset (Nguyen et al., 2025) is used as the training set for the VLSP 2025 DRiLL task on legal document retrieval. The dataset contains 3,129 expert-annotated triplets of the form $\{question, relevant\ articles, answer\}$ and 59,636 articles across 27 domains of Vietnamese law.

The preprocessing pipeline transforms raw legal documents into a structured corpus for retrieval. As shown in Figure 2, the steps include:

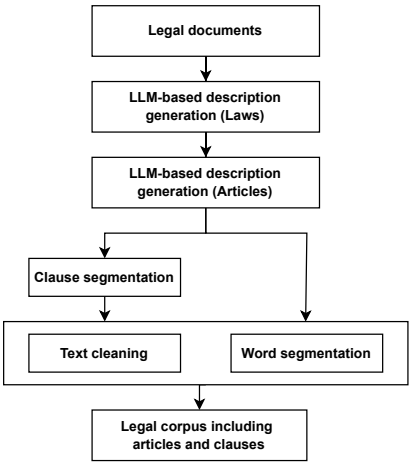

Figure 2: Preprocessing pipeline for legal documents.

1. **Description generation:** Summaries are generated at law and article levels using a large language model, specifically Qwen2.5-72B-Instruct (Team et al., 2024).

2. **Clause segmentation:** Splitting articles into smaller clauses for fine-grained search.

3. **Text normalization:** Cleaning whitespace, punctuation, and redundant symbols.

4. **Word segmentation:** Applying Vietnamese tokenization tools such as VnCoreNLP (Vu et al., 2018).

The output is a normalized, hierarchical corpus of laws, articles, and clauses enriched with descriptions.

3.3 Candidate Retrieval

At this stage, up to 500 relevant legal articles are extracted from a corpus of nearly 60,000 articles. The objective of this stage is to maximize recall while substantially reducing the number of articles considered for subsequent re-ranking. To achieve this, a hybrid search mechanism is employed, combining lexical matching and semantic embeddings, and further enhanced through the use of HyDE to expand the semantic scope of the query.

The procedure consists of three steps:

1. **Candidate retrieval using the original query:** Hybrid search is applied to generate an initial ranked list of candidates based on a hybrid score, which combines BM25 scores and embedding similarity. The list is traversed sequentially from the top; for each retrieved article, if references to other articles are present, those referenced articles are incorporated into the candidate list, contributing up to 20% of the target size, corresponding to at most 80 additional articles. This process continues until 400 candidate articles are obtained.

2. **Candidate retrieval using the HyDE-generated query:** The HyDE method is utilized to generate a semantically enriched hypothetical query derived from the original question. Hybrid search is then performed in the same manner as step 1, producing another set of 400 candidate articles.

3. **Merging of candidate sets:** The two candidate sets from steps 1 and 2 are merged and re-ranked based on the hybrid score. The top 500 articles with the highest scores are selected as the final candidate pool for the subsequent re-ranking stage.

This approach leverages both the original and HyDE-generated queries while exploiting the inter-article reference structure, allowing the retrieval scope to be expanded without degrading the overall quality of the results.

3.4 Multi-Stage Ranking

After obtaining the set of 500 candidate legal articles from the Candidate Retrieval stage, a multi-stage re-ranking strategy is applied to progressively refine and prioritize the most relevant documents. The overall objective of this stage is to improve ranking precision while preserving high recall, ensuring that the subsequent reasoning modules operate on a compact yet highly informative subset of legal texts. The re-ranking process consists of three main stages, each leveraging different levels of semantic and structural signals.

Feature	Description
reranker	Relevance score from BGE v2 M3 neural reranker.
search	Semantic similarity using BGE-M3 embeddings.
normalized	Retrieval score normalized to reduce scale variance.
HyDE	Embedding distance from HyDE-generated query representation.
cos_VVP, cos_DVT	Cosine similarity using two domain-specific models (VoVanPhuc, DangVanTuan).
rrf	Reciprocal Rank Fusion score with pseudo-queries.
dot, Euclidean, L1	Vector similarity and distance metrics for diversity.
article_idx, clause_idx	Article index and clause index in the corpus.
is_law, is_clause	Binary indicators for full law or individual clause.
doc_length_token, length_diff, length_ratio, query_length_token, length_diff_token, length_ratio_token	Length-related features capturing token counts, length ratios and differences.
sentence_count, comma_count	Structural indicators representing stylistic and punctuation characteristics of legal texts.

Table 1: Feature set used for LambdaMART-based re-ranking in Stage 1.

Stage 1: From Top 500 to Top 100. In the first stage, a pairwise learning-to-rank framework based on the LambdaMART algorithm is utilized to re-order the initial 500 candidates retrieved in the previous step. LambdaMART, a gradient-boosted decision tree model optimized for ranking tasks, is particularly suitable for handling heterogeneous features and capturing complex interactions between them. The model is trained on a comprehensive and diverse set of features that encode various semantic, lexical, and structural aspects of the legal documents, including but not limited to:

The inclusion of heterogeneous features allows the LambdaMART model to capture both semantic relevance and document-level structural characteristics. Furthermore, the model's hyperparameters, such as the number of trees, learning rate, and maximum tree depth, are tuned automatically using Optuna (Akiba et al., 2019). This automated optimization ensures that the re-ranker achieves optimal performance across multiple evaluation metrics, such as Normalized Discounted Cumulative Gain (NDCG) and Mean Reciprocal Rank (MRR), without manual trial-and-error.

Feature	Description
reranker_score	Baseline relevance score from the BGE v2 M3 neural reranker.
ltr_score	LambdaMART ranking score produced in Stage 1, aggregating pairwise learning signals.
llm_score	Semantic alignment score computed by feeding each query–article pair into the LLMs, returning a continuous relevance score in $[0, 1]$.
rrf_score	Reciprocal Rank Fusion score combining the original and HyDE-generated query signals.
normalized_scores	Normalized versions of the above features (e.g., normalized reranker, LTR, LLM, and RRF scores) to reduce distributional bias and ensure comparability across heterogeneous scoring mechanisms.

Table 2: Feature set used for LightGBM ranker-based listwise re-ranking in Stage 2.

Stage 2: From Top 100 to Top 10 In the second re-ranking stage, we apply a more fine-grained listwise learning-to-rank strategy to refine the 100 candidates returned by Stage 1. We use a LightGBM ranker with `objective=rank_xendcg` and `metric=ndcg`. This listwise objective directly optimizes the quality of the ranked list as measured by NDCG. Unlike the pairwise optimization used in Stage 1, this listwise approach optimizes the quality of the entire ranked list, enabling the model to capture global ranking consistency among the remaining highly relevant candidates.

Similar to the first stage, the LightGBM ranker operates on a set of semantically enriched features derived from multiple sources, including LLM evaluations. Table 2 summarizes the key features utilized in Stage 2.

By leveraging both learned pairwise ranking signals from Stage 1 and context-sensitive semantic evaluations from the LLM, Stage 2 provides a more nuanced assessment of legal article relevance. The use of `xendcg` as the optimization objective allows the model to directly maximize metrics aligned with downstream evaluation criteria such as NDCG.

Hyperparameter Optimization. To ensure optimal performance of the LightGBM ranker, we employ Optuna, a state-of-the-art automated hyperparameter optimization framework. Key parameters such as the number of boosting rounds, learning rate, maximum tree depth, and feature subsampling ratios are tuned using Optuna's Bayesian optimiza-

tion and pruning strategies. This automated search process enables efficient exploration of the hyperparameter space, yielding configurations that maximize ranking metrics (e.g., NDCG@10 and MRR) without extensive manual trial-and-error.

This final refinement yields a highly precise and semantically coherent top-10 ranked list of legal articles. The two-stage design, which combines pairwise and listwise learning paradigms, has been empirically demonstrated to outperform single-stage approaches by exploiting complementary ranking signals at different levels of granularity.

3.5 Answer Selection

After obtaining the top-10 candidates from the multi-stage re-ranking process, the final step is to precisely identify the legal provisions that are directly relevant to the input question. To achieve this, we adopt a pointwise selection framework that integrates LTR features with the contextual reasoning capabilities of LLMs.

Specifically, the process consists of three main components:

1. **Leveraging LTR features from the previous stage:** The re-ranking scores produced by LambdaMART are used as input features for the pointwise model. The goal is to exploit the ranking signals accumulated in the earlier stage as a foundation for evaluating the relevance of each candidate.

2. **Incorporating LLM with Chain-of-Thought:** For each $\{question, legal\_provision\}$ pair, the LLM is prompted to conduct a step-by-step analysis in which the key legal elements in the question (entities, referenced statutes, and the legal relationship at issue) are extracted, these elements are aligned with the content of the candidate provision, and a continuous relevance score is produced. This CoT-based evaluation mitigates the risk of overlooking critical provisions, particularly in cases involving indirect references or complex linguistic expressions.

3. **Collective evaluation and score fusion:** Beyond pairwise assessment, the full set of 10 candidate provisions is presented to the LLM together with the question, and collective chain-of-thought reasoning is requested. This enables the model to capture cross-provision relationships such as conditions, exceptions, and mutual citations, thereby improving the selection of the final answer set. The resulting scores from collective reasoning, together with those from pairwise evaluation and LTR features, are integrated into a pointwise LightGBM model. All features are normalized using RobustScaler, and the model outputs relevance probabilities for each candidate. Instead of fixing the number of top-k results, an optimal threshold determined on the validation set based on the F2-score is applied to balance coverage and precision.

By combining these three components, the *Answer Selection* stage goes beyond relying solely on ranking features and incorporates deep contextual reasoning from LLMs. This ensures that the selected legal provisions are not only highly accurate but also supported by transparent reasoning, thereby laying a robust foundation for subsequent legal answer generation.

4 Experimental Results Analysis

4.1 Computational Resources

All experiments were conducted on a single AI server equipped with 4× NVIDIA L40 GPUs, each with 48 GB of VRAM and a total of 192 GB of VRAM, along with 256 GB of system RAM and 96 CPU cores. Training of the LambdaMART and LightGBM models required approximately 10 minutes, while preprocessing of the VLQA dataset took 24 hours.

To evaluate response time, we tested the complete pipeline on a dataset of 219 queries. The average execution time for each stage was as follows: Stage Candidate Retrieval and Stage 1 of Multi-Stage Ranking for initial retrieval and filtering took 2.1 minutes, Stage 2 of Multi-Stage Ranking for reranking with LLM-based scoring took 45.6 minutes, and Stage Answer Selection for final answer selection took 23.2 minutes. In total, the pipeline required approximately 70.9 minutes to process the entire test set, which corresponds to an average response time of about 19.4 seconds per query across all stages.

4.2 Candidate Retrieval

In the first stage, our primary objective is to maximize recall, ensuring that the majority of relevant legal provisions are included in the candidate pool for subsequent re-ranking. This step is particularly

critical in the legal domain, where missing a relevant statute could lead to incorrect reasoning or incomplete legal justification.

Method	P@500	R@500
BM25	0.0025	0.9408
Dense retrieval	0.0026	0.9712
Hybrid search	0.0026	0.9732
Hybrid + mentioned	0.0026	0.9741
Hybrid + mentioned + HyDE	**0.0026**	**0.9833**

Table 3: Candidate retrieval performance on the VLSP 2025 DRiLL training set.

As shown in Table 3, all retrieval methods achieve high recall, thereby minimizing the risk of missing relevant statutes. Although BM25 reaches a relatively strong recall of 0.9408, it still fails to capture a considerable proportion of relevant documents. Dense retrieval improves recall to 0.9712, highlighting the advantage of semantic representation over keyword-based approaches.

When lexical and semantic signals are combined in Hybrid search, recall further improves to 0.9732, showing the benefits of leveraging both approaches. More notably, augmenting candidate retrieval with inter-article references (Hybrid + mentioned) yields a small but consistent gain, with R@500 of 0.9741, confirming the importance of citation structures in legal texts. Finally, the integration of HyDE achieves the highest recall of 0.9833, demonstrating the usefulness of context generation in expanding retrieval coverage. Importantly, selecting the top 500 candidates provided recall above 0.98 while still keeping the subsequent re-ranking process computationally manageable.

4.3 Multi-Stage Ranking

Stage 1: From Top 500 to Top 100. The first re-ranking stage aims to refine the initial candidate set and enhance precision. Table 4 compares different methods in this stage. Selecting the top-100 solely based on candidate retrieval scores maintains recall at 0.9408. In contrast, re-ranking with models such as LightGBM-binary and rerankers significantly improves recall, with LightGBM-binary combined with a reranker reaching 0.958.

The LambdaMART model with feature normalization provides further gains. In particular, the RobustScaler configuration achieves the best results, with R@100 of 0.9655 and P@100 of 0.0124. These results demonstrate that pairwise learning-to-rank effectively exploits semantic and structural features to improve top-100 coverage.

Method	P@100	R@100
Top 100 Candidate Retrieval	0.0120	0.9408
LightGBM-binary without reranker	0.0122	0.9499
Reranker	0.0122	0.9546
LightGBM-binary with reranker	0.0123	0.9580
LambdaMART with StandardScaler	0.0123	0.9605
LambdaMART with RobustScaler	**0.0124**	**0.9655**

Table 4: Re-ranking performance from top 500 to top 100 on the VLSP 2025 DRiLL private test set.

Stage 2: From Top 100 to Top 10. In the final refinement stage, a listwise LightGBM ranker is employed; as shown in Table 5, it outperforms the Stage 1 baseline, raising Recall from 0.8662 to 0.8803 and F2 from 0.3587 to 0.3636.

Method	P@10	R@10	F2
Stage 1 + Reranker	0.1070	0.8628	0.3576
Only Stage 1	0.1072	0.8662	0.3587
Stage 1 + LightGBM Ranker	**0.1086**	**0.8803**	**0.3636**

Table 5: Re-ranking performance from top 100 to top 10 on the VLSP 2025 DRiLL private test set.

These stages ensure a high recall rate while substantially reducing the volume of data for subsequent stages.

4.4 Answer Selection

The results in Table 6 highlight the critical role of the Answer Selection stage in improving both precision and recall compared with relying solely on multi-stage re-ranking. Specifically, when LLM-based scores are incorporated across all candidates in the configuration *Multi-Stage Ranking + LLM all*, recall increases from 0.7110 to 0.7565, while F2 rises from 0.6714 to 0.7038. These results indicate that leveraging the contextual reasoning capability of LLMs enables the system to capture relevant legal provisions more effectively.

More importantly, when step-by-step reasoning via CoT is further integrated with Multi-Stage Ranking and LLM applied to all candidates, precision improves substantially from 0.5502 to 0.6773, while recall remains high at 0.7394. Consequently, F2 reaches its highest value of 0.7261, indicating that CoT not only enhances accuracy but also helps maintain a balanced trade-off between coverage and reliability.

Methods	Precision	Recall	F2
Multi-Stage Ranking	0.5492	0.7110	0.6714
Multi-Stage Ranking + LLM all	0.5502	**0.7565**	0.7038
Multi-Stage Ranking + LLM all + CoT	**0.6773**	0.7394	**0.7261**

Table 6: Answer Selection performance on the VLSP 2025 DRiLL private test set.

In summary, the Answer Selection stage demonstrates the effectiveness of combining LTR-derived features with LLM-based contextual reasoning, particularly when augmented with CoT. This integration provides a robust foundation for the subsequent stage of legal answer generation.

4.5 Error Analysis

Although the system achieves competitive results, several errors remain. Ambiguous queries and underspecified legal concepts often caused retrieval failures, while citation expansion sometimes added irrelevant statutes that reduced precision. In reranking, the LLM sometimes prioritized surface word overlap over deeper relevance, and limited use due to computation led to both false negatives and false positives.

Category	Percent (%)	Precision	Recall	F2
what	41.56	0.679	0.724	0.714
how	24.66	0.536	0.910	0.799
yes/no	21.46	0.478	0.803	0.707
when	4.57	0.783	0.800	0.797
which	4.11	0.420	0.722	0.632
who	3.65	0.635	0.938	0.856
where	0.45	0.500	0.500	0.500

Table 7: Performance by question category on the evaluation set. Percent indicates the proportion of each category.

Table 7 summarizes results by question type. Categories such as *who* and *when* achieved balanced precision and recall, while binary and enumerative *which* and *where* questions showed higher error rates due to ambiguity. Overall, these findings suggest that future improvements should focus on query reformulation, selective citation handling, and more efficient semantic scoring.

Limitations

While the proposed multi-stage pipeline shows promising performance on Vietnamese legal retrieval, several directions remain for further investigation. First, pretraining dense encoders on Vietnamese statutes and case law would likely enhance their ability to capture legal semantics. Second, combining LLM-based reasoning with symbolic methods or rule-based checks could provide stronger support, particularly for handling statutory exceptions and cross-references. Third, system interpretability remains an open challenge: although Chain-of-Thought reasoning provides partial transparency, future work should explore structured explanation frameworks for legal decision-making. Finally, broader evaluations beyond the VLSP 2025 DRiLL dataset are needed to assess the system's robustness and adaptability across diverse Vietnamese legal corpora.

Conclusion

This paper presented a system for the Vietnamese legal document retrieval task in the VLSP 2025 DRiLL Challenge. We proposed a multi-stage pipeline that integrates lexical retrieval, semantic embeddings, re-ranking, and prompting with large language models, while enriching signals through generated titles and inter-article references. Within a learning-to-rank framework, these components were progressively combined to refine candidate rankings, achieving state-of-the-art performance in the competition. Our results confirm the effectiveness of hybrid retrieval strategies and feature enrichment for Vietnamese legal texts, providing strong evidence that multi-stage architectures can balance recall and precision in low-resource legal NLP. Future research will explore more advanced re-ranking techniques, deeper integration of domain knowledge, and extensions to other tasks, thereby contributing to the broader development of legal NLP in low-resource settings.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81.

Ilias Chalkidis, Manos Fergadiotis, Ion Androutsopoulos, and Nikolaos Aletras. 2020. Legal-bert: The

muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Juliano Rabelo, Ken Sato, Hiroshi Kiyomaru, Lam Minh Thang Nguyen, Minh Le Tran, Jack Reese, Frederik Moller, and Ricardo Almeida. 2021. Coliee-2021: Overview of the 8th competition on legal information extraction and entailment. In *Proceedings of the Eighth International Competition on Legal Information Extraction/Entailment (COLIEE 2021)*.

Vladimir Karpukhin, Barlas Oğuz, 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.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Yun Ma, Guido Zuccon, Andy Nguyen, and Gianluca Demartini. 2020. Easing legal search: Listwise learning-to-rank with bert for precedent and news retrieval. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1501–1504.

Ha-Thanh Nguyen, Thi-Hai-Yen Nguyen, Minh-Binh Tran, Minh-Le Nguyen, and Tuan-Anh Dang. 2023. Vietlegal: A large-scale vietnamese legal dataset for legal information retrieval. *arXiv preprint arXiv:2305.12345*.

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. *Preprint*, arXiv:2507.19995.

Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. In *Foundations and Trends® in Information Retrieval*, volume 3, pages 333–389.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, and ... 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*. Version v2, revised on 3 Jan 2025.

Van-Hoang Tran, Thi-Thanh Nguyen, and Minh-Tien Le. 2022. Applying learning-to-rank for query autocompletion and sentence ranking in vietnamese legal documents. *arXiv preprint arXiv:2204.05678*.

Xuan-Son Vu, Minh Le Nguyen, Viet Dac Nguyen, and Hong Phuong Le. 2018. Vncorenlp: A vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 56–60. ACL.

Thi-Hai-Yen Vuong, Tan-Minh Nguyen, Hoang-Trung Nguyen, Trong-Khoi Dao, and Le Hoang-Quynh Nguyen, Ha-Thanh. 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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Qiang Wu, Christopher JC Burges, Krysta M Svore, and Jianfeng Gao. 2010. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270.

Haoxi Zhong, Chaojun Xiao, Zhipeng Tu, Zhiyuan Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12169*.