DRILL Shared Task 2025: The Challenge of Deep Retrieval in the Expansive Legal Landscape

Thi-Hai-Yen Vuong*[1], Tan-Minh Nguyen[2], Hoang-Trung Nguyen[1], Trong-Khoi Dao[3], Ha-Thanh Nguyen[4], Hoang-Quynh Le[1]

[1]VNU University of Engineering and Technology, Hanoi, Vietnam
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[3]VNU University of Law, Hanoi, Vietnam
[4]National Institute of Informatics, Tokyo, Japan

Abstract

This paper presents a summary of the DRILL Shared Task, which focuses on legal information retrieval in the Vietnamese language as part of the Vietnamese Language and Speech Processing workshop. Over the two-month competition, more than 50 teams participated in developing retrieval systems to address legal queries across diverse domains. Most teams follow the common retrieve-then-rerank paradigm, with a fine-grained processing step at the end. Nonetheless, the difference in top-performing teams is that they develop a learning-to-rank model from various features, including large language models (LLMs). These meticulously designed approaches achieved strong performance, surpassing the baseline methods. Despite these gains, the inherent complexity of legal texts indicates significant opportunities for further advancement.

1 Introduction

The legal system governs a wide range of everyday human activities, such as civil rights, finance, education, family and marriage, thereby shaping the lives of both legal professionals and laypeople (Ponce et al., 2019). Recent advances in artificial intelligence (AI), particularly in natural language processing (NLP), have opened new opportunities for legal applications to address real-world needs. The increasing digitization of legal texts, combined with the capabilities of modern AI models, provides a strong foundation for narrowing the gap between legal expertise and public understanding (Zhong et al., 2020). While legal NLP research in languages such as English, Chinese, and Japanese is already well established, Vietnamese remains comparatively underexplored. To address this gap, we introduce, through the Vietnamese Language and Speech Processing (VLSP) workshop, a foundational shared task in the Vietnamese legal domain[1].

This research aims to advance the development of retrieval systems for juridical documents, a core problem in legal NLP. In general, retrieval systems identify documents relevant to a given query from a large corpus. In the legal context, a document retrieval task focuses on locating statutory articles that are relevant to or supportive of a given legal question within a collection of legal codes and statutes.

However, the legal document retrieval task presents two major challenges. The first is the scarcity of resources required to develop reliable retrieval systems, particularly the limited availability of high-quality annotated data, which often demands expert domain knowledge. The second is the inherent complexity of accurately mapping user queries to relevant statutory articles. This challenge arises from the unique characteristics of legal texts, including hierarchical structures, specialized terminology, and intricate cross-referencing among legal provisions. As a result, general semantic matching techniques may fall short in capturing the nuanced relationships necessary for effective legal retrieval.

To address these challenges, the Challenge on Deep Retrieval in the expansive Legal Landscape (DRILL) is organized to encourage new research in legal NLP studies, particularly for low-resource languages. The DRILL benchmark contains more than 3,000 legal issues raised by Vietnamese citizens and refined by experts to ensure quality and reliability. Participants were tasked with retrieving the correct relevant articles from a large database of 59,636 legal documents. To the best of our knowledge, this work is one of the first large-scale, high-quality practical Vietnamese datasets tailored for legal information retrieval.

In this paper, we present a comprehensive overview of the DRILL shared task, providing a detailed description of the task definition, the underlying VLQA dataset, developed systems, and the evaluation framework. We also discuss the

*Corresponding author
[1]https://vlsp.org.vn/vlsp2025/eval/drill

significance of this initiative for advancing Vietnamese legal NLP research and its potential impact on developing accessible legal assistance tools for Vietnamese speakers. The shared task represents one of the first major initiatives aimed at advancing Vietnamese Legal NLP, addressing the growing need for intelligent legal text processing applications in low-resource language settings.

2 DRILL Benchmark

2.1 Task Definition

The DRILL shared task focuses on Legal Document Retrieval, a fundamental challenge in legal NLP. Given a question q and a corpus $A = \{a_1, a_2, ..., a_m\}$, this task aims to learn a retrieval model that returns $A' \subset A$ where each article $a_i \in A'$ is considered "relevant" to the corresponding question q. We define an article as "relevant" to a query if the query sentence can be answered Yes/No, or can be entailed from the meaning of the article. This definition ensures that the retrieved articles provide sufficient information to address the legal question posed.

2.2 Data Construction

The DRILL benchmark was constructed from the VLQA dataset (Nguyen et al., 2025), which contains law issues posed by Vietnamese citizens on public legal consultation platforms. To ensure quality and relevance, an initial automated filtering step was applied to remove overly short or irrelevant posts. This was followed by a more refined manual filtering process to preserve the diversity and richness of the corpus. Annotation was carried out iteratively by five senior law students and a legal expert across multiple rounds until the data satisfied all criteria. A comprehensive annotation guideline, including detailed instructions and examples, was provided to standardize the process. Any discrepancies among annotators were carefully reviewed and resolved through discussion to maintain consistency.

Figure 1 provides a detailed example of the data format used for training, illustrating how each legal question is associated with relevant articles, as well as the structure of the law corpus containing article contents.

2.3 Data Statistics

Table 1 summarizes statistics on dataset size, question length, and the number of relevant articles per


```
[
  {
    "qid": 11938,
    "question": "Chế độ báo cáo của doanh nghiệp kinh
doanh dịch vụ xếp hạng tín nhiệm quy định thế nào?",
    "relevant_laws": [
      27053,
      27071
    ]
  }
]
```


(Training data)


```
[
  {
    "id": 0,
    "law_id": "14/2022/TT-NHNN",
    "content": [
      {
        "aid": 0,
        "content_Article": "Điều 1. Phạm vi điều chỉnh"
      },
      {
        "aid": 1,
        "content_Article": "Điều 2. Chức danh ..."
      },
    ]
  }
]
```


(Provided article corpus)

Figure 1: Data format used for training. The top box shows annotated training samples with question ID, legal question, and corresponding relevant law article IDs. The bottom box shows the structure of the law corpus, where each law contains multiple articles identified by article ID and associated legal content.

query across the subsets. The data are divided into three parts: a training set with 2,190 samples, a public test set with 312 samples, and a private test set with 627 samples. To mitigate overfitting and reduce bias in the public test, the private test set was designed to be twice as large.

An inherent complexity of DRILL benchmark is that a single query may correspond to multiple relevant articles. On average, each query maps to 1.34 articles, with a maximum of nine.

We further categorize the dataset into five domains: Economics and Finance (EF), State Management and Law (SL), Society, Culture, and Education (SCE), Infrastructure and Development (ID), and Science and Technology (ST). Domain-specific statistics are reported in Table 2. The EF and SL domains together account for 67.11% of the dataset, reflecting the most common legal challenges faced by laypeople. By contrast, only about 7% of questions fall within the ST domain, primarily concerning technology and intellectual property.

Table 1: Statistics of the DRILL data.

	Train	Public test	Private test
# samples	2190	312	627
# words per question			
Average	19.71	20	19.90
Minimum	6	11	11
Maximum	45	42	44
# relevant articles per question			
Average	1.34	1.31	1.32
Minimum	1	1	1
Maximum	9	4	5

2.4 Competition Framework

The competition is hosted on the Codabench platform[2], a standardized platform for organizing machine learning benchmarks and shared tasks. The competition consists of several consecutive phases:

- **Registration Phase**: Participants register and form teams.
- **Development Phase**: Training data is released for participants to develop and fine-tune their models.
- **Evaluation Phase**: Public and private test sets are used to assess model performance. Participants may submit results multiple times per day, with the public leaderboard available to support iterative development.
- **Submission Phase**: Participants submit final system outputs along with source code and a brief system description.
- **Results and Publication Phase**: Final rankings based on the private test set are released. Selected teams are invited to present their systems at the shared task session of the conference.

To encourage iterative refinement, the leaderboard during the evaluation phase reflects performance on a public subset of the test set. The private test set is released only shortly before the submission deadline, ensuring fairness in the final evaluation.

2.5 Data Usage Restrictions

To ensure fairness and reproducibility, the competition imposes the following restrictions:

- **External Data**: Participants are not allowed to use any external data in any part of the processing pipeline.
- **Pre-trained Models**: Only models that were publicly released before the year the competition is held are permitted. The use of closed-source or proprietary language models (e.g., GPT-4o, Gemini) is strictly prohibited.
- **Reproducibility**: Each submission must include sufficient information to reproduce the results, including instructions for accessing or reconstructing the models used.

To ensure fair enforcement of the competition's policies across all participants, each team is required to submit a brief report outlining their proposed solution, the pre-trained models and LLMs employed, and the corresponding reproducible implementation. We only accept results from participants who correctly follow our guidelines to prevent any violations of the competition objectives.

3 System Descriptions and Performance

The competition was hosted on Codabench (Xu et al., 2022), an online platform for organizing AI benchmarks and challenges. During the evaluation phase, each team can submit at most 10 times per day. The leaderboard is also visible to participants, allowing them to refine their methods based on the results obtained on the public test set. The private test set is provided to participants only three days before the submission deadline, after submitting the source code and a brief system description. The organizers would verify that the submissions are reproducible using the submitted source code. The official evaluation metrics are recall, precision, and macro-average F2 scores.

$$Recall = \text{avg}\frac{\text{\# correctly retrieved articles per query}}{\text{\# relevant articles per query}}$$

$$Precision = \text{avg}\frac{\text{\# correctly retrieved articles per query}}{\text{\# retrieved articles per query}}$$

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

3.1 Baseline systems

Following prior studies in IR (Robertson and Zaragoza, 2009; Trotman et al., 2014; Rosa et al., 2021), we employ two baseline models: a statistical ranking model and a two-stage retrieval pipeline.

[2] https://www.codabench.org/competitions/9722/

Table 2: Data distribution by domain.

Topic	Description	Train	Public	Private	Total
Economics and Finance	business, commerce, investment, etc.,	737	106	205	1048
State management and Law	administrative violations, civil rights, criminal liability, etc.	743	102	207	1052
Society, Culture and Education	labor & wages, culture, education, healthcare, etc.,	265	39	75	379
Infrastructure and Development	real estate, natural resources & environment, etc.	297	42	92	431
Science and Technology	technology, intellectual property, etc.,	148	23	48	219

BM25 is a scoring algorithm that computes the relevance of documents given a query based on term frequencies. This model achieves an F2 score of 0.3365 on the VLQA private test set. The latter pipeline involves leveraging a cross-encoder architecture as a further re-ranking step, which is illustrated in Figure 2. The cross-encoder model was fine-tuned using negative articles derived from BM25 outputs and positive (relevant) articles following the cross-entropy loss function. The ratio of negative articles to positive articles is $5:1$. The two-stage baseline achieves a 0.5512 F2 score on the private test set, showing an improvement by 12 points over the BM25 baseline.

3.2 Participants approaches

DRILL shared task has attracted more than 50 participants, with a total of 1,222 submissions for three phases: public test, private test, and post challenge. In this section, we present an overview of the approaches taken by the teams that submitted papers describing their methods.

EDM developed a four-phase pipeline involving various techniques: the first phase employs a hybrid search using Hypothetical Decoument Embedding (HyDE) to produce pseudo query embeddings to retrieve candidate articles. The second phase reduces the search space via pair-wise learning to rank using input features from the retrieval scores from the previous phase, re-ranker scores, cosine similarity computed by embedders, and statistical characteristics. Next, they further curate candidate articles using evaluation from LLM (Qwen2.5-32B, Qwen2.5-72B, LLaMA3.3-70B). Finally, they combined all features from previous steps via a point-wise learning to rank model to return top$-k$ output articles.

Unknown first augmented the data by generating the headers and titles of legal articles using zero-shot prompting and a Qwen2.5-7B model. After that, they proposed an approach based on a combination of a re-ranker (BGE-reranker-v2-m3) and LLM (Qwen2.5-72B) using the boosting ensemble approach.

ducanger pre-processed the article corpus by collecting the title of the article from external sources. They then applied a two-stage approach involving a combination of BM25 and BGE-m3, a text embedding model as the initial retrieval step, followed by a re-ranking step based on a fine-tuned BGE-reranker-v2-m3 with hard negative samples mined from the previous step. Finally, a filtering score process is performed to filter out candidates based on a pre-defined threshold.

fasterunited developed a three-phase pipeline following a retrieval-then-rerank paradigm. First, lengthy articles are split into smaller chunks. They then employ BM25, followed by fine-tuned embedding models (E5-Instruct, GTE-multilingual) to retrieve candidate articles. Next, a BGE-rerank model is fine-tuned with contrastive loss and hard negative samples. Finally, relevant articles are determined based on a pre-defined threshold.

truong13012004 also followed a retrieval-then-rerank paradigm with a combination of BM25 and fine-tuned BGE-reranker-v2-m3 model. They first fine-tuned a Vietnamese text embedding model (Duc et al., 2024) and then performed a semantic search to take the top-300 candidate articles. Next, a fine-tuned re-ranker and BM25 are employed to calculate the similarity between the question and candidates. Finally, scores from BM25 and re-ranker are combined under multiple weight settings.

ngjabach proposed a four-step approach involving fine-tuning bi-encoder, cross-encoder models, and an LLM-based voting mechanism. They first retrieve candidate articles using GTE-multilingual-E2. These articles are then re-ranked using Vi-Ranker (Phuong, 2024) to determine the top-10 most relevant articles. Finally, Qwen2.5-14B and Qwen2.5-32B are employed to perform a multiple-round voting and "debating" ensemble.

Engineers first employed a combination of BM25 and fine-tuned multilingual-E5-large to retrieval top-100 candidate articles. They then fine-tuned a re-ranking model using contrastive loss and hard negative articles from the previous step.

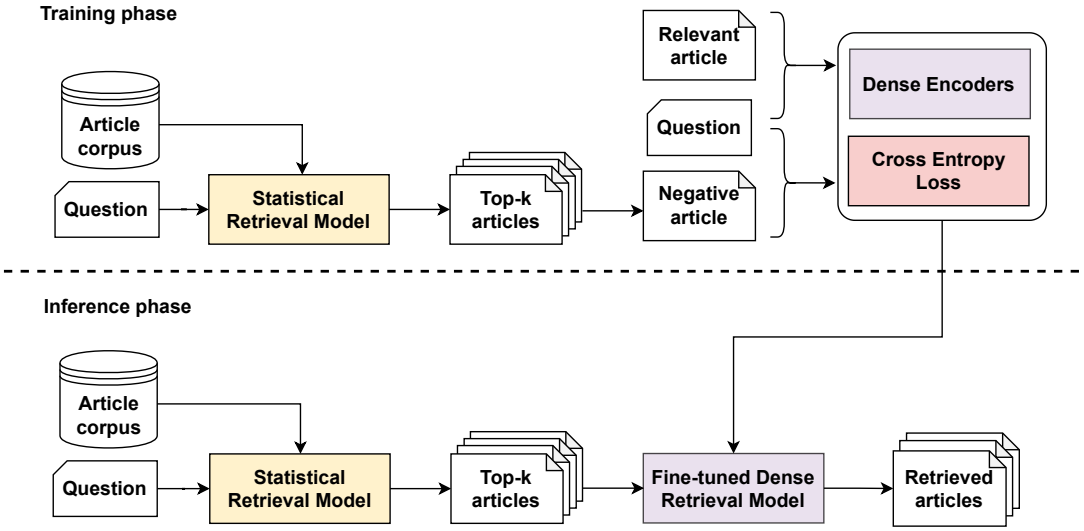

Figure 2: Overall architecture of the two-stage baseline model. Candidates are first retrieved from the document corpus using BM25. Next, a cross-encoder model is fine-tuned using negative documents derived from BM25 outputs and positive (relevant) documents following the cross-entropy loss.

Table 3: Top-10 participants and baselines on the public test leaderboard. The best and second-best results are highlighted in boldface and underlined, respectively.

#	Participant	F2	Precision	Recall
1	unknow_123	**0.7426**	**0.5810**	**0.7981**
2	ducanger	<u>0.7345</u>	0.5748	<u>0.7893</u>
3	truong13012004	0.7032	0.5563	0.7529
4	villageai	0.6704	0.5197	0.7228
5	AImba	0.6473	<u>0.5759</u>	0.668
6	fasterunited	0.6337	0.4907	0.6835
7	edmmm	0.5982	0.4144	0.6728
8	SoftMind_AIO	0.5905	0.3649	0.6985
9	ngjabach	0.5812	0.3391	0.7075
10	nguyentai090301	0.5758	0.3841	0.6579
24	BM25 Baseline	0.2771	0.2767	0.2772

Table 4: Top-10 participants and baselines on the private test leaderboard. † denotes teams that show a performance improvement compared to Table 3.

#	Participant	F2	Precision	Recall
1	edmmm†	**0.7261**	**0.6773**	0.7394
2	unknow_123	<u>0.6966</u>	<u>0.6222</u>	0.7181
3	ducanger	0.6955	0.5097	**0.7653**
4	dinhanhx	0.6710	0.5509	0.7097
5	fasterunited†	0.6521	0.4153	<u>0.7605</u>
6	truong13012004	0.6495	0.4714	0.7172
7	AImba	0.6425	0.4086	0.7498
8	ngjabach	0.6280	0.4329	0.7077
9	Engineers	0.5864	0.3147	0.7478
10	villageai	0.5587	0.3799	0.6332
12	2-stage Baseline	0.5512	0.3740	0.6253
19	BM25 Baseline	0.3365	0.2265	0.3830

Articles whose re-ranking score is greater than a pre-defined threshold are selected as outputs.

4 Deeper Analysis

Tables 3 and 4 present the results of the top 10 participants and the baselines on the public and private leaderboards, respectively. Most teams outperformed the baselines, typically by employing traditional IR techniques such as BM25, text embedding models for semantic similarity, more recent LLMs, or hybrid approaches that combined them. A common pipeline emerged: statistical or bi-encoder models were used in the initial retrieval stage to gather candidate articles, followed by more computationally intensive methods such as re-ranking models or LLMs to refine results on a smaller set of articles. Notably, 8 of the top 10 teams on the public leaderboard experienced performance declines on the private test set, with the exceptions being fasterunited and edmmm. Their robustness may be attributed to careful preprocessing strategies and the integration of features from multiple models. A deeper analysis of participant performance reveals that top-ranking teams consistently leveraged advanced retrieval techniques, including HyDE, learning-to-rank, and LLM-based rerankers. Furthermore, data-centric strategies, such as data augmentation, summarization, and decomposition, also played a critical role in the strong results achieved by the top three teams.

We further analyze the performance of the top

Table 5: F2 performance of top-10 participants in the private test set by domain.

Participant	EF	SL	SCE	ID	ST
edmmm	0.6820	**0.7407**	**0.7402**	**0.7775**	0.6942
unknow_123	0.6843	0.7047	0.6455	0.7316	0.7061
ducanger	**0.6966**	0.6667	0.6700	0.7415	**0.7439**
dinhanhx	0.6915	0.6590	0.6059	0.6932	0.6928
fasterunited	0.6485	0.6412	0.6384	0.6854	0.6834
truong13012004	0.6906	0.6202	0.5928	0.6678	0.6412
AImba	0.6592	0.6297	0.5833	0.6728	0.6768
ngjabach	0.6617	0.5964	0.5678	0.6549	0.6640
Engineers	0.5982	0.5493	0.5631	0.6155	0.6566
villageai	0.5345	0.5607	0.5481	0.6187	0.5310
Average	0.6547	0.6369	0.6155	0.6859	0.6690

teams in the private test set across domains as presented in Table 5. Edmmm stands out as the most consistent top performer, ranking first in State management and Law (SL), Society, Culture and Education (SCE), and Infrastructure and Development (ID). Ducanger leads in Economics and Finance (EF) and Science and Technology (ST), showing particular strength in finance reasoning and science fields, highlighting differences in generalization capability across teams. SCE is the most challenging domain for participants, except edmmm, due to its interdisciplinary nature. These results reveal a gap in domain adaptation, indicating room for improvement in the future.

5 Conclusion and Future work

The DRILL Shared Task demonstrated the potential of NLP techniques in tackling text information retrieval across diverse legal domains. Within just two months, from the call for participation to the completion of the private test phase, we received more than 1,100 submissions from over 50 teams. These numbers highlight both the community's strong interest and the significance of this research problem. The top-performing teams employed combinations of LLMs and fine-tuned pretrained language models, achieving improvements over the baselines, though often by modest margins. Notably, a substantial performance drop was observed in the Social, Culture, and Education domain, showing promising areas for further improvements. In conclusion, we are pleased to report that the DRILL Shared Task was a success and is well-positioned to make meaningful contributions to the Vietnamese legal NLP community.

Limitations

As our work aims to provide researchers with a well-defined benchmark for evaluating existing methods and developing robust models for legal information retrieval, certain limitations should be acknowledged.

First, although the DRILL benchmark covers legal questions from various domains, it is limited to Vietnamese legislation. Extending the study to other languages would further support the development of more robust and advanced legal information retrieval methods.

Second, not all legal questions can be addressed using articles alone. Further operations, such as information extraction and answer generation, are necessary to meet the demands of real-world legal support systems. Future work will therefore focus on building high-quality, reliable question-answering benchmarks for legal applications.

Ethics Statement

We strive to adhere to the ACL Code of Ethics. Our technology is intended to supplement, not replace, legal professionals, with careful attention to responsible use and awareness of potential limitations and biases in automated systems. All data used in this research have been anonymized and stripped of personally identifiable information in accordance with relevant data protection regulations. The datasets employed are derived from publicly available online sources and do not infringe upon the proprietary rights of any individuals or entities.

Acknowledgments

We would like to express our sincere gratitude to the VLSP organizers for hosting the workshop that enabled us to collectively address the DRILL challenge, to the sponsors and supporters for their essential contributions in making the event possible, and to all participants for their valuable efforts, which play a crucial role in advancing research in this field.

References

Nguyen Quang Duc, Le Hai Son, Nguyen Duc Nhan, Nguyen Dich Nhat Minh, Le Thanh Huong, and Dinh Viet Sang. 2024. Towards comprehensive vietnamese retrieval-augmented generation and large language models. *arXiv preprint arXiv:2403.01616*.

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. *arXiv preprint arXiv:2507.19995*.

Nam Dang Phuong. 2024. Viranker: A cross-encoder model for vietnamese text ranking.

Alejandro Ponce, Sarah Chamness Long, Elizabeth Andersen, Camilo Gutierrez Patino, Matthew Harman, Jorge A Morales, Ted Piccone, Natalia Rodriguez Cajamarca, Adriana Stephan, Kirssy Gonzalez, and 1 others. 2019. Global insights on access to justice 2019: Findings from the world justice project general population poll in 101 countries. *World Justice Project*, page 1.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Guilherme Moraes Rosa, Ruan Chaves Rodrigues, Roberto Lotufo, and Rodrigo Nogueira. 2021. Yes, bm25 is a strong baseline for legal case retrieval. *arXiv preprint arXiv:2105.05686*.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, ADCS '14, page 58–65, New York, NY, USA. Association for Computing Machinery.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.