VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

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*

Received 02nd December 2025
Revised 23rd December 2025; Accepted 18th May 2026

Abstract: This paper provides an overview of the DRILL Shared Task, a Vietnamese legal information retrieval challenge organized under the Vietnamese Language and Speech Processing workshop. Over the two-month competition period, more than 50 teams participated, contributing a total of 1,255 submissions to the leaderboard. While most teams adopted a standard retrieve-then-rerank pipeline complemented by a final fine-grained processing stage, the top-performing teams distinguished themselves by constructing learning-to-rank models enriched with diverse features, including those derived from large language models (LLMs). These carefully engineered methods delivered strong results, outperforming baseline systems. However, our error analysis reveals that current systems struggle with questions involving commonsense knowledge, extremely long context, and temporal relations, suggesting avenues for future work.

Keywords: Vietnamese NLP, Legal Information Retrieval, Article Retrieval, Shared Task, Deep Learning.

1. Introduction

The legal system shapes numerous aspects of everyday life—including civil rights, finance, education, and family affairs—affecting both legal practitioners and the general public [1]. Recent progress in artificial intelligence (AI), particularly in natural language processing, has opened up new avenues for developing legal applications that address practical, real-world needs. With the growing availability of digitized legal texts and the increasing power of modern AI models, there is a strong foundation for reducing the gap between legal expertise and public understanding [2]. Although legal NLP has seen substantial development in languages such

*Corresponding author.
E-mail address: lhquynh@vnu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.6463

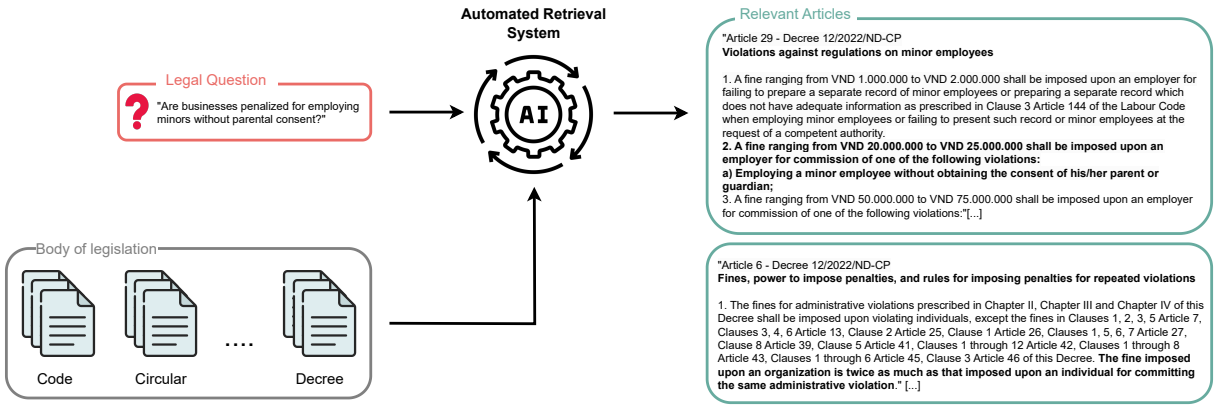

Figure 1. Illustration of statutory article retrieval task in DRILL. All examples we show in the paper are translated from Vietnamese for illustration.

as English, Chinese, and Japanese, Vietnamese research remains relatively limited. To help bridge this gap, we introduce a foundational shared task in the Vietnamese legal domain through the Vietnamese Language and Speech Processing (VLSP) workshop[1].

This work seeks to advance retrieval methods for statutory documents, a fundamental task in legal NLP. Retrieval systems typically aim to locate documents in a large corpus that are relevant to a user's query; in the legal setting, this involves identifying statutory articles that address or support a given query, as shown in Figure 1.

Yet legal document retrieval poses two central challenges. The first is the lack of resources needed to build robust systems, most notably the shortage of high-quality annotated data, which often requires domain experts to produce. The second challenge lies in the difficulty of accurately linking user queries to the correct statutory articles. Legal texts contain hierarchical structures, domain-specific terminology, and dense cross-references, making them particularly hard to interpret. Consequently, general semantic matching techniques frequently fall short of capturing the nuanced relationships required for effective legal retrieval.

In this work, we propose Deep Retrieval in the expansive Legal Landscape (DRILL) dataset to foster progress in legal NLP, particularly for low-resource languages. The DRILL benchmark comprises more than 3,000 legal questions posed by Vietnamese citizens and refined by domain experts to ensure accuracy and reliability. Participants were required to retrieve the appropriate statutory articles from a large repository of 59,636 legal documents. To the best of our knowledge, this represents one of the first large-scale, human-annotated Vietnamese datasets designed specifically for legal information retrieval.

Furthermore, we provide a thorough overview of the DRILL shared task in VLSP 2026, detailing the task formulation, benchmark construction, participating systems, and evaluation methodology. We further discuss the broader significance of this effort for Vietnamese legal NLP and its potential to support the development of accessible legal-assistance technologies for Vietnamese speakers. As one of the earliest major initiatives in Vietnamese Legal NLP, the shared task directly addresses the growing demand for intelligent legal text-processing tools in low-resource language contexts.

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

To conclude, the main contributions of this work are threefold:

- **DRiLL benchmark:** We present DRiLL, a novel dataset comprising over 3,000 legal issues raised by citizens spanning across various domains. To the best of our knowledge, this work is one of the first large-scale, human-annotated datasets for Vietnamese legal information retrieval.
- **VLSP 2025 Shared Task:** We organize the challenge of Deep Retrieval in the Expansive Legal Landscape at VLSP 2025. This initiative encouraged community participation and benchmarking, fostering progress in legal information processing for the Vietnamese.
- **Comprehensive analyses:** We present in-depth studies on the performance of baselines and participant approaches to offer valuable insights for further improvements. Our findings reveal promising directions for handling implicit reasoning, long-context documents, and cross-article dependencies in legal texts.

2. Related Work

Classical information retrieval models largely rely on lexical overlap between queries and documents, while early neural approaches [3, 4] enhanced retrieval by capturing contextual relevance. More recently, the advent of LLMs such as GPT-4 and Gemini has facilitated retrieval-augmented generation (RAG) [5–7], in which retrieved documents serve as external knowledge to improve the performance in domain-specific tasks.

In recent years, the research community has developed several benchmark datasets to support the advancement of robust models for a variety of legal NLP tasks. For example, CJRC is a Chinese judicial reading comprehension dataset comprising over 50,000 question–answer pairs derived from the factual descriptions of 10,000 legal cases [8]. CAIL2018, introduced by [9], represents the first large-scale Chinese dataset for legal judgment prediction, encompassing more than 2.6 million criminal cases associated with approximately 400 legislative articles and criminal charges. Similarly, [10] compiled a dataset for legal judgment prediction containing 11,532 admissible cases from the European Court of Human Rights (ECtHR). In addition, [11] proposed the first class action legal judgment prediction dataset, consisting of 5,459 lost cases and 5,290 won cases. Beyond judgment prediction, [12] developed CAIL2019-SCM, which comprises roughly 9,000 case triplets designed for case similarity detection. For legal summarization tasks, [13] introduced RulingBR, a Portuguese dataset containing approximately 10,000 rulings from the Brazilian Federal Supreme Court. In a related effort, [14] constructed a benchmark for summarizing legal contracts into plain English, including 446 contract action sets paired with corresponding reference summaries.

Statutory article retrieval has received relatively limited attention within the legal NLP community, primarily due to the scarcity of large-scale, high-quality datasets. Existing resources include the COLIEE Statute Law corpus [15], BSARD [16], and ALQAC [17]. The COLIEE dataset comprises 1,206 questions from the Japanese bar examination, each annotated with references to the relevant provisions of the Japanese Civil Code. Similarly, ALQAC provides 100 question–article pairs curated by legal experts and jurists. However, both datasets are centered on bar exam questions, which differ substantially from the queries posed by laypersons in real-world contexts: bar exam questions are highly technical and domain-specific, whereas everyday legal inquiries are generally broader and more straightforward. More recently, [16] introduced BSARD, a French

statutory retrieval dataset containing 1,108 legal questions linked to 22,633 statutory articles, sourced from legal consulting organizations. In a similar vein, DRILL also draws its questions from actual citizen inquiries but distinguishes itself through its larger scale and broader coverage, encompassing 3,129 annotated questions and nearly 60,000 Vietnamese statutory articles. This expansive scope provides a robust foundation for the development of practical, data-driven legal information retrieval systems.

3. DRILL Shared Task

3.1. Task Definition

The DRILL shared task centers on Legal Article Retrieval, a core problem in legal NLP. Given a legal question q and a corpus of statutory articles $A = \{a_1, a_2, ..., a_m\}$, the objective is to develop a retrieval model that identifies a subset $A' \subset A$ in which each article $a_i \in A'$. is *relevant* to q. We consider an article relevant if its content can either answer the question in a Yes/No manner or if its meaning entails the query. This criterion ensures that retrieved articles provide adequate information to address the legal issue at hand.

The complexity of the legal domain necessitates a departure from standard retrieval metrics. In legal retrieval, the main requirement is to avoid missing any article that may affect the legal interpretation of a query. A false negative is usually more harmful than a false positive, because missing a relevant legal basis can lead to an incorrect or incomplete legal conclusion. Metrics such as F1, MAP, or nDCG treat precision and recall more evenly or focus mainly on ranking quality, which does not fully reflect this recall-heavy requirement.

For this reason, DRILL adopts the F2-score, which weights Recall twice as much as Precision. This choice encourages models to retrieve a broader set of potentially relevant articles and better matches the real needs of legal analysis, where completeness of retrieved legal grounds is more important than strict precision at top ranks.

Furthermore, the legal basis required to address a query fully is rarely contained within a single article. Many legal queries are linked to multiple relevant articles rather than a single one. These articles often play different roles: one may define general rules, another provides exceptions, while others specify scope, conditions, or procedures. Therefore, the retrieval task becomes multi-label, where the model must return a set of articles instead of one best match. This challenge is increased by cross-document dependencies. A single article may not fully answer a query on its own, and the correct interpretation only emerges when several articles are read together. As a result, the model must recognize not only text similarity but also relationships between articles across the corpus. These factors make legal retrieval in DRILL harder than standard retrieval tasks, since the model must capture structural and semantic connections across multiple documents and ensure high recall to support complete legal reasoning.

3.2. Data Construction

The DRILL benchmark is derived from the VLQA dataset [18], which collects questions raised by Vietnamese citizens on public legal consultation platforms. The annotation proceeded through multiple rounds on a Streamlit-based website. There are three roles in the annotation process: *manager*, *annotator*, and *expert*. The *manager* is directly responsible for operating and maintaining the annotation tool. The *annotator* verifies the collected dataset and makes changes if needed. Then the *expert* reviews and classifies them into five categories:

1. Satisfies all the requirements
2. The annotated articles are invalid
3. The annotated articles are inactive.
4. The annotated articles are incomplete.
5. The annotated articles are redundant.

Table 1. Agreement measurement between experts and annotators

Category	Count	Per.
Satisfies all the requirements	2738	87.5%
The annotated articles are invalid	151	4.8%
The annotated articles are incomplete.	91	2.9%
The annotated articles are redundant.	81	2.6%
The annotated articles are inactive.	68	2.2%

All annotators and experts are native speakers who have been carefully trained with detailed annotation guidelines. Specifically, five senior law students are serving as annotators. Two PhDs are experts in Vietnamese Law and NLP. Table 1 reports the measurement of agreement between experts and annotators. More than 87% of annotated samples demonstrate full agreement, where all quality criteria are satisfied. The other cases require further discussion between annotators and the expert.

Figure 2 illustrates the annotated data format, showing how each legal question is linked to its relevant articles and how the statutory corpus is structured. Specifically, the training data format is organized into two sections. The top section displays annotated samples, including the question ID, the legal question itself, and the IDs of relevant law articles. The bottom section presents the structure of the law corpus, where each law comprises multiple articles, each identified by an article ID and accompanied by its associated legal content.

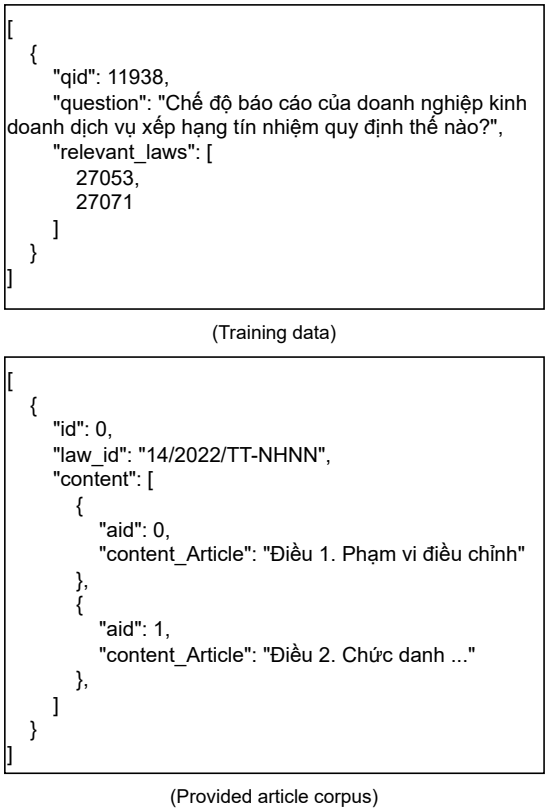

Figure 2. Illustration of samples and article corpus.

3.3. Data Statistics

Table 2 provides an overview of the dataset's characteristics, including its size, average question length, and the distribution of relevant articles per query across the three data splits. The benchmark is divided into 2,190 training examples, 312 public test examples, and 627 private test examples, with the private test set intentionally made larger to help limit overfitting and reduce evaluation bias. A notable aspect of the DRILL benchmark is that a single query may correspond to multiple statutory articles. On average, each question is linked to 1.34 relevant articles, with some queries requiring up to nine.

The dataset is further organized 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-level statistics are provided in Table 3. The EF and SL domains together constitute 67.11% of the dataset, highlighting the most frequently encountered legal issues among the general public. In contrast, the ST domain accounts for only about 7% of questions, primarily addressing topics related to technology and intellectual property.

Table 4 classifies legal reasoning in the dataset into four categories: lexical matching (LM), semantic interpretation (SI), logical inference (LI), and multi-article reading

Table 2. 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

(MAR). Semantic interpretation constitutes the largest portion (59%), reflecting questions that demand comprehension of legal semantics, including synonyms, antonyms, and nuanced interpretations. Lexical matching, by contrast, relies primarily on surface-level similarity and direct phrase alignment without external context. This type is comparatively easier for retrieval, especially for models leveraging keyword matching or retrieval-augmented approaches. Logical inference, representing roughly one-fifth of the dataset, involves binary reasoning such as Yes/No questions or conditional statements, requiring models to both identify relevant content and perform logical deductions. Finally, multi-article reading, accounting for about 25% of questions, is the most demanding, necessitating the integration of information across multiple statutory articles, including cross-referencing and temporal reasoning, such as handling repealed laws.

3.4. Competition Framework

The DRILL shared task was hosted on the Codabench platform[2], a standardized environment for organizing machine learning benchmarks and competitions. The competition proceeded through several sequential phases:

- **Registration Phase**: Participants register and form teams.
- **Development Phase**: Training data is provided to develop and fine-tune models.
- **Evaluation Phase**: Both public and private test sets are used to evaluate performance. Participants can submit multiple times per day, with the public leaderboard reflecting a subset of the test set to support iterative improvements.
- **Submission Phase**: Final system outputs, source code, and brief descriptions are submitted.
- **Results and Publication Phase**: Final rankings are determined using the private test set, and selected teams are invited to present their systems during the shared task session at the conference.

To promote iterative refinement while ensuring fairness, the leaderboard only reflects performance on the public subset during the evaluation phase, with the private test set released shortly before the submission deadline.

3.5. Data Usage Restrictions

To maintain fairness and reproducibility, the competition enforces the following restrictions:

- **External Data**: Participants are prohibited from using any external data at any stage of the processing pipeline.
- **Pre-trained Models**: Only publicly released models available before the competition year are allowed. The use of closed-source or proprietary language models, such as GPT-4o or Gemini, is strictly forbidden.
- **Reproducibility**: Submissions must include sufficient information to reproduce the results, including instructions for accessing or reconstructing the models used.

To ensure compliance, each team is required to submit a concise report detailing their proposed

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

Table 3. Distribution and brief description of data domain in DRILL

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

Table 4. Distribution and examples of different reasoning types of questions in DRILL. One question may require multiple reasoning abilities, so the sum of percentages is over 100%

Type	Per.	Examples
Lexical Matching	41%	**Question**: How is business name registration regulated? **Article 18**. Business name registration [...]
Semantic Interpretation	59%	**Question**: What documents are required to import veterinary drug samples for registration? **Article 22**. Registration for import of veterinary drugs and veterinary drug ingredients [...]
Logical Inference	22%	**Question**:Can I use other documents to carry out birth registration procedures for my child? **Article 16**. Birth registration procedures [...]
Multi-Article Reading	25%	**Question**: Are defense workers eligible for salary increases when their salary is transferred? **Article 3**. Salary regime for defense workers [...] **Article 4**. Salary transfer 1. Principles of salary transfer [...]

solution, the pre-trained models and LLMs employed, and the corresponding reproducible implementation. Only results from participants adhering to these guidelines are accepted, safeguarding the integrity and objectives of the competition.

4. System Descriptions and Performance

The competition was hosted on Codabench [19], an online platform for organizing AI benchmarks and challenges. During the evaluation phase, each team was allowed up to 10 submissions per day. The public leaderboard provided visibility into performance on the public test set, enabling participants to iteratively refine their models. The private test set was released only three days before the submission deadline, once teams had submitted their source code along with a brief system description. Organizers verified the reproducibility of all submissions using the provided code. The official evaluation metrics for the competition were recall, precision, and the macro-averaged F2 score.

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

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

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

4.1. Baseline systems

One-stage retriever Following established approaches in information retrieval [20–22], we implement two baseline models: a statistical ranking model and a two-stage retrieval pipeline. BM25, a classic term-frequency-based scoring algorithm, ranks documents according to their relevance to a given query. This model achieves an F2 score of 0.3365 on the private test set.

Table 5. Performance of BM25 Baseline, 2-stage Baseline, and ablations. Top-k denotes returning the top-k retrieved articles

Method	F2	Precision	Recall
BM25 Baseline	0.3368	0.2265	0.3835
2-stage Baseline	0.5515	0.3740	0.6257
Ablations of 2-stage Baseline			
w/o BM25 negatives	0.4691	0.1883	**0.7479**
top-1 retrieval	0.5496	**0.6316**	0.5323
top-2 retrieval	**0.5515**	0.3740	0.6257
top-5 retrieval	0.4615	0.1825	0.7473

Two-stage retriever The second baseline employs a two-stage pipeline that incorporates a cross-encoder re-ranking step, as illustrated in Figure 3. The cross-encoder is fine-tuned using relevant (positive) articles and negative articles sampled from BM25 outputs, with a negative-to-positive ratio of 5:1, optimized using the cross-entropy loss. This two-stage approach achieves an F2 score of 0.5512 on the private test set, representing an improvement of approximately 0.215 over the BM25 baseline.

Beyond the primary baselines, a series of ablation experiments is conducted to examine the contribution of individual components within the two-stage architecture. As shown in Table 5, removing BM25-derived negative samples produces a clear decline in overall F_2 performance (from 0.5515 to 0.4691). This configuration yields substantially higher recall but markedly lower precision, suggesting that negative examples mined from BM25 outputs play a crucial role in enabling the cross-encoder to distinguish relevant from non-relevant articles. In their absence, the model tends to over-generalize and retrieve a broader set of irrelevant candidates.

The impact of varying the number of retrieved candidates is also investigated. Restricting the system to top-1 retrieval leads to a slight reduction in F_2, as the configuration favours precision at the expense of recall. Expanding the pool to top-2 candidates restores the optimal balance between the two metrics and matches the performance of the full baseline. By contrast, increasing the pool to top-5 introduces considerably more non-relevant documents, thereby reducing precision and lowering the overall F_2 score. These observations indicate that the effectiveness of the two-stage pipeline is sensitive to both the quality and the size of the BM25 candidate set. A compact, well-curated top-k retrieval list consistently provides more stable and effective supervision than a larger and noisier candidate pool.

4.2. *Participants approaches*

The DRILL shared task attracted over 50 teams, resulting in 1,222 submissions across the public test, private test, and post-challenge phases. Here, we summarize the strategies of teams that submitted detailed descriptions of their approaches:

EDM implemented a four-phase pipeline: starting with hybrid retrieval using Hypothetical Document Embeddings (HyDE) to generate pseudo query embeddings, followed by pair-wise learning-to-rank using features from previous retrieval scores, re-ranker outputs, cosine similarities, and statistical characteristics. Candidate articles were further curated with evaluations from LLMs (Qwen2.5-32B, Qwen2.5-72B, LLaMA3.3-70B), and a final point-wise learning-to-rank model integrated all features to select the top-k articles.

FPT IS augmented the data by generating article headers and titles via zero-shot prompting with Qwen2.5-7B. Their approach combined a re-ranker (BGE-reranker-v2-m3) and LLM (Qwen2.5-72B) using a boosting ensemble.

CUDO pre-processed article titles from external sources, then applied a two-stage pipeline: initial retrieval with BM25 and BGE-m3 embeddings, followed by fine-tuning of BGE-reranker-v2-m3 with hard negatives, and a final filtering step based on a predefined threshold.

IUH.TD followed a three-phase retrieval-then-rerank pipeline: splitting long articles

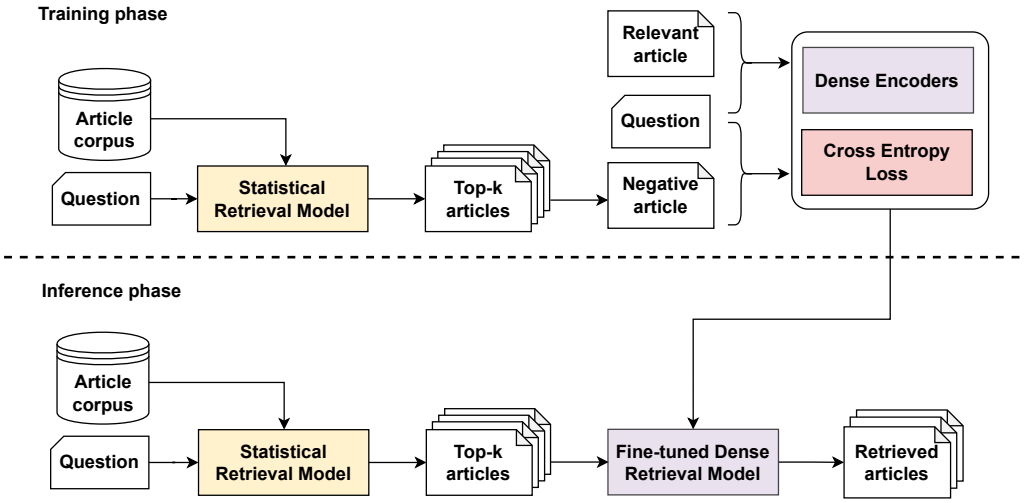

Figure 3. Overall architecture of the two-stage baseline model. Candidates are first retrieved from the document corpus using BM25. Next, a cross-encoder model is fine-tuned using negative documents derived from BM25 outputs and positive (relevant) documents following the cross-entropy loss.

into chunks, retrieving candidates with BM25 and embedding models (E5-Instruct, GTE-multilingual), then re-ranking with a fine-tuned BGE-reranker using contrastive loss and hard negatives, with final selection via thresholding.

brain_rot also used a retrieval-then-rerank approach: top-300 candidates were obtained via semantic search with fine-tuned Vietnamese embeddings [23], then re-ranked using BM25 and a fine-tuned BGE-reranker, with final scores combined under multiple weighting schemes.

BK_DRILL proposed a four-step pipeline: initial candidate retrieval using GTE-multilingual-E2, re-ranking with ViRanker [24] to obtain top-10 articles, and multi-round voting and "debating" ensembles using Qwen2.5-14B and Qwen2.5-32B.

Engineers retrieved top-100 candidates using BM25 combined with multilingual-E5-large embeddings, then fine-tuned a re-ranking model with contrastive loss and hard negatives, selecting articles exceeding a predefined score threshold as final outputs.

Table 6. 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	FPT IS	**0.7426**	**0.5810**	**0.7981**
2	CUDO	<u>0.7345</u>	0.5748	<u>0.7893</u>
3	brain_rot	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	IUH.TD	0.6337	0.4907	0.6835
7	EDM	0.5982	0.4144	0.6728
8	Engineers	0.5931	0.4057	0.6753
9	BK_DRiLL	0.5812	0.3391	0.7075
10	nguyentai090301	0.5758	0.3841	0.6579
24	BM25 Baseline	0.2771	0.2767	0.2772

5. Main results

Tables 6 and 7 summarize the results of the top 10 teams alongside the baseline models on the public and private leaderboards. Most participants exceeded baseline performance by leveraging a combination of traditional information retrieval methods (e.g., BM25), text embedding models for semantic similarity, modern LLMs, or hybrid pipelines that integrated these approaches. A common strategy involved

Table 7. Top-10 participants and baselines on the private test leaderboard. † denotes teams that show a performance improvement

#	Participant	F2	Precision	Recall
1	EDM	**0.7261**	**0.6773**	0.7394
2	FPT IS	0.6966	0.6222	0.7181
3	CUDO	0.6955	0.5097	**0.7653**
4	Berry	0.6710	0.5509	0.7097
5	IUH.TD	0.6521	0.4153	0.7605
6	brain_rot	0.6495	0.4714	0.7172
7	AImba	0.6425	0.4086	0.7498
8	BK_DRiLL	0.6280	0.4329	0.7077
9	Engineers	0.5864	0.3147	0.7478
10	villageai	0.5587	0.3799	0.6332
11	NaiveNotNice	0.5578	0.2769	0.7472
12	2-stage Baseline	0.5512	0.3740	0.6253
13	Bosch@AI Team	0.5496	0.3712	0.6247
14	Stepbystep	0.4861	0.3301	0.5512
15	IUH-Strix	0.4671	0.3150	0.5313
16	DAAH	0.4423	0.2869	0.5116
17	Mât Vụ Black	0.3834	0.286	0.4191
18	Whisperers	0.3379	0.1337	0.5469
19	BM25 Baseline	0.3365	0.2265	0.3830

using statistical or bi-encoder models for initial candidate retrieval, followed by computationally intensive re-ranking or LLM-based refinement on the reduced set of articles.

The leading team, *EDM*, develops a multi-stage pipeline that integrates features from multiple sources (statistical characteristics, cosine similarity, dense retriever scores, LLM-as-judge). Their solution serves as a filter, where the irrelevant articles are removed iteratively after each stage. As a result, the *EDM* team achieves a strong precision of 0.6773, surpassing the top-2 team by 5% point, while maintaining a relatively good recall score.

Interestingly, 8 of the top 10 teams on the public leaderboard show declines on the private test set, with the exceptions being *IUH.TD* and *EDM*. Their robustness is likely due to careful pre-processing and the integration of features from multiple models. For example, the *IUH.TD* team applies structure-aware chunking to split lengthy legal articles into smaller segments, enabling comprehensive representation

Table 8. F2 performance of top-10 participants in the private test set by domain

Participant	EF	SL	SCE	ID	ST
EDM	0.6820	**0.7407**	**0.7402**	**0.7775**	0.6942
FPT IS	0.6843	0.7047	0.6455	0.7316	0.7061
CUDO	**0.6966**	0.6667	0.6700	0.7415	**0.7439**
Berry	0.6915	0.6590	0.6059	0.6932	0.6928
IUH.TD	0.6485	0.6412	0.6384	0.6854	0.6834
brain_rot	0.6906	0.6202	0.5928	0.6678	0.6412
AImba	0.6592	0.6297	0.5833	0.6728	0.6768
BK_DRiLL	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

of articles while preserving essential legal structure and domain knowledge. Furthermore, data-centric strategies, including augmentation, summarization, and segmentation, play a crucial role in the solutions of the three leading teams.

6. Additional analysis

6.1. Performance by domains

We further examine top-team performance across domains on the private test set, as shown in Table 8. The *Society, Culture, and Education* domain proves the most challenging for most teams, likely due to its close connection to everyday activities. Therefore, the question would be more complex, requiring logical reasoning based on multiple articles. In contrast, the easiest domain, *Infrastructure and Development*, focuses on how a specific statement is regulated in the legislation, which can be identified by matching overlapped words between the question and articles. These findings reveal differences in domain generalization among teams and point to potential areas for future improvement in domain adaptation strategies.

6.2. Performance by reasoning type

Participants' results are further examined by categorizing questions according to the type of reasoning required, as shown in Table 9. Overall, multi-article reading, which requires integrating information across multiple legal provisions,

Table 9. F2 performance of top-10 participants in the private test set by the reasoning type of question

Participant	LM	SI	LI	MAR
EDM	**0.7146**	**0.6875**	**0.7087**	**0.5419**
FPT IS	0.6705	0.6584	0.5982	0.4594
CUDO	0.6928	0.5967	0.5468	0.5051
Berry	0.6482	0.6149	0.5690	0.4684
IUH.TD	0.5060	0.4811	0.4594	0.4066
brain_rot	0.6314	0.5713	0.5521	0.4904
AImba	0.5897	0.5702	0.5199	0.4976
BK_DRiLL	0.5962	0.5538	0.5082	0.4285
Engineers	0.5235	0.4586	0.4410	0.4235
villageai	0.5397	0.5022	0.4983	0.4100
Average	0.5801	0.5423	0.5190	0.4408

emerges as the most challenging category, showing F2 scores roughly 10 percentage points lower than the other types. In contrast, lexical matching questions are the easiest, as they contain substantial word overlap between the query and returned documents, allowing retrieval systems to identify relevant articles more effectively. Notably, EDM achieves the best performance across all reasoning types, with a substantial margin over all other participants. This consistent advantage underscores the strength of its multi-stage framework, which integrates advanced techniques such as HyDE, LLM-as-reranker, and point-wise learning-to-rank to capture both surface-level and deeper semantic signals.

6.3. Error analysis

Among the 627 samples in the private test set, 30 questions are considered easy, achieving an average F2 score above 90. In contrast, 8 questions are so challenging that no participating system successfully retrieves the relevant articles. Several representative hard cases are presented in Table 10.

One such example is Question 641, which requires commonsense knowledge that the national emblem appears on Land Use Rights certificates. Because the question text does not explicitly mention the term "national emblem," automated retrieval systems instead rely on surface-level keywords such as "Sổ hồng" or "red envelope," leading them to return irrelevant documents related to forgery or providing false information. This reveals a critical weakness: current retrieval models struggle when essential semantic links are not explicitly stated in the query.

Another difficult case is Question 3088, which challenges systems with extremely long statutory documents. The relevant article spans 34 clauses and over one thousand words, with the decisive information appearing near the end. Models relying on fixed-length truncation or sparse representations fail to capture the final clause, resulting in incorrect retrieval. This underscores the need for more sophisticated long-context processing strategies for legal documents.

Finally, cross-document dependencies and temporal relations pose substantial obstacles. For instance, Article 15 of Circular 17/2020/ND-CP amends Article 8 of Circular 87/2018/ND-CP. Most participants' systems incorrectly return Article 8 rather than Article 15, the amended version, because they do not track amendment chains or legal evolution over time. This limitation illustrates the broader challenge of understanding complex inter-document relationships within statutory frameworks.

These observations demonstrate that handling implicit knowledge, long-context statutory structure, and cross-document dependencies remains beyond the capability of most current approaches. Building a reference graph over legal documents, incorporating amendment relations, hierarchical structure, and semantic dependencies, could offer a promising future direction for addressing such complex retrieval scenarios.

7. Discussions

Our analysis reveals several valuable insights underlying system performance in the Vietnamese statutory retrieval task. The top-performing systems consistently combine

Table 10. The most challenging samples that no system can correctly retrieve relevant articles

ID	Question	Relevant articles
641	Can printing the image of a Land Use Rights Certificate (Sổ hồng) on red envelopes lead to criminal liability?	**Article 351** - Criminal Code "Desecration of national flag, national emblem, national anthem"
3088	Is a business manager also the owner of a sole proprietorship?	**Article 4** - Law on Enterprises [...] 24. "executive of an enterprise means the owner of a sole proprietorship, a general partner of a partnership, chairperson or member of the Member/Partner Assembly, President of a company, President or member of the Board of Directors, Director/General Director, or holder of another managerial position prescribed in the company's charter."
13478	What are the conditions for traders in the business of buying and selling liquefied petroleum gas?	**Article 8** - Circular 87/2018/ND-CP "Requirements applied to gas sellers and purchasers. 1. Every gas seller and purchaser must: a) be a trader in accordance with provisions of laws; b) own gas tanks meeting safety requirements or LPG bottles eligible for circulation on the market or enter into a contract of tank or LPG bottle lease;[...]" **Article 15** - Circular 17/2020/ND-CP "Amendments to some Articles of the Government's Decree No. 87/2018/ND-CP dated June 15, 2018 on gas business" "2. Point b Clause 1 of Article 8 is amended as follows:[...]"

traditional sparse retrieval (BM25, Elasticsearch), dense neural embeddings, and LLM-based re-ranking or reasoning modules. Each component plays a distinct role: sparse retrieval ensures broad initial coverage, dense retrieval improves semantic matching, and LLMs refine candidate rankings by capturing deeper contextual relations. This multi-stage design proves far more effective than any single retrieval paradigm. Top teams frequently employ careful pre-processing steps, including document segmentation, clause-level indexing, and collecting articles' titles. These techniques help reduce noise and compress long documents, thereby improving the overall retrieval performance.

Despite promising performance on lexical matching and well-aligned queries, systems struggle with implicit reasoning, temporal relations, and cross-document dependencies. This indicates that automated retrieval systems have not fully comprehended domain-specific legal reasoning patterns in Vietnamese law. The limitation underscores the need for dedicated legal corpora and methods to model amendment chains, hierarchical structure, and multi-article logic. Our findings and error analyses reveal the limitations of the current LLM-based approach in practical legal scenarios. The developed 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.

8. Conclusion and Future work

The DRILL Shared Task highlighted how NLP methods can effectively support text retrieval across a wide range of legal domains. In two months, from the call for participation to the end of the private test phase, we received over 1,100 submissions from more than 50 teams, underscoring both the community's enthusiasm

and the importance of this research challenge. The strongest systems combined LLMs with fine-tuned pre-trained models and consistently outperformed the baselines. However, our error analysis reveals that commonsense knowledge, extremely long context, and temporal relations are the main bottlenecks of the current systems, suggesting avenues for future work. We are pleased to conclude that the DRILL Shared Task was organized successfully, offering a solid foundation for further developments in legal NLP for Vietnamese law.

Acknowledgments

This work has been supported by VNU University of Engineering and Technology under project number CN25.12.

We extend our heartfelt thanks to the VLSP organizers for providing the workshop platform that enabled our joint efforts on the DRILL challenge, to the sponsors and supporters whose contributions made the event possible, and to all participants whose dedication continues to drive progress in this research area.

Hoang-Trung Nguyen was supported by the Program "Innovation and Quality Enhancement in Postgraduate Training" of the University of Engineering and Technology, Vietnam National University, Hanoi.

References

[1] A. Ponce, S. C. Long, E. Andersen, C. G. Patino, M. Harman, J. A. Morales, T. Piccone, N. R. Cajamarca, A. Stephan, K. Gonzalez, et al., Global Insights on Access to Justice 2019: Findings from the World Justice Project General Population Poll in 101 Countries, World Justice Project2019, pp. 1.
URL https://worldjusticeproject.org/sites/default/files/documents/WJP-A2J-2019.pdf

[2] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, M. Sun, How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence, D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5218–5230. https://doi.org/10.18653/v1/2020.acl-main.466.
URL https://aclanthology.org/2020.acl-main.466/

[3] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense Passage Retrieval for Open-Domain Question Answering, B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550.
URL https://aclanthology.org/2020.emnlp-main.550/

[4] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, J. Lin, End-to-End Open-Domain Question Answering with BERTserini, W. Ammar, A. Louis, N. Mostafazadeh (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 72–77. https://doi.org/10.18653/v1/N19-4013.
URL https://aclanthology.org/N19-4013/

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented Generation for Knowledge-intensive NLP Tasks, Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.
URL https://dl.acm.org/doi/abs/10.5555/3495724.3496517

[6] B. J. Gutiérrez, Y. Shu, Y. Gu, M. Yasunaga, Y. Su, HippoRAG: Neurobiologically Inspired Long-term Memory for Large Language Models, Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24, Curran Associates Inc., Red Hook, NY, USA, 2024.
URL https://dl.acm.org/doi/10.5555/3737916.3739818

[7] C. Nguyen, P. Nguyen, L.-M. Nguyen, Retrieve–Revise–Refine: A novel framework for retrieval of concise entailing legal article set, Information Processing Management, Vol. 62, No. 1, 2025, pp. 103949. https://doi.org/10.1016/j.ipm.2024.103949.
URL https://www.sciencedirect.com/science/article/pii/S030645732400308X

[8] X. Duan, B. Wang, Z. Wang, W. Ma, Y. Cui, D. Wu, S. Wang, T. Liu, T. Huo, Z. Hu, et al., CJRC: A Reliable Human-Annotated Benchmark DataSet for Chinese Judicial Reading Comprehension, Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18, Springer, 2019, pp. 439–451.

URL https://dl.acm.org/doi/abs/10.1007/978-3-030-32381-3_36

[9] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, J. Xu, CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction, ArXiv, Vol. abs/1807.02478, 2018, https://doi.org/10.48550/arXiv.1807.02478.

[10] M. Medvedeva, M. Vols, M. Wieling, Judicial decisions of the European Court of Human Rights: Looking into the crystal ball, Proceedings of the conference on empirical legal studies, 2018, p. 24.

[11] G. Semo, D. Bernsohn, B. Hagag, G. Hayat, J. Niklaus, ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US, N. Aletras, I. Chalkidis, L. Barrett, C. Goanță, D. Preoțiuc-Pietro (Eds.), Proceedings of the Natural Legal Language Processing Workshop 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 31–46. https://doi.org/10.18653/v1/2022.nllp-1.3.
URL https://aclanthology.org/2022.nllp-1.3/

[12] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, T. Zhang, X. Han, Z. Hu, H. Wang, J. Xu, CAIL2019-SCM: A Dataset of Similar Case Matching in Legal Domain, ArXiv, Vol. abs/1911.08962, 2019, https://doi.org/10.48550/arXiv.1911.08962.

[13] D. de Vargas Feijó, V. P. Moreira, RulingBR: A Summarization Dataset for Legal Texts, Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2018, p. 255–264, https://doi.org/10.1007/978-3-319-99722-$3_2$6.

[14] L. Manor, J. J. Li, Plain English Summarization of Contracts, N. Aletras, E. Ash, L. Barrett, D. Chen, A. Meyers, D. Preotiuc-Pietro, D. Rosenberg, A. Stent (Eds.), Proceedings of the Natural Legal Language Processing Workshop 2019, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1–11. https://doi.org/10.18653/v1/W19-2201.
URL https://aclanthology.org/W19-2201/

[15] R. Goebel, Y. Kano, M.-Y. Kim, J. Rabelo, K. Satoh, M. Yoshioka, Overview of Benchmark Datasets and Methods for the Legal Information Extraction/Entailment Competition (COLIEE) 2024, New Frontiers in Artificial Intelligence: JSAI International Symposium on Artificial Intelligence, JSAI-IsAI 2024, Hamamatsu, Japan, May 28–29, 2024, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2024, p. 109–124, https://doi.org/10.1007/978-981-97-3076-6_8.

[16] A. Louis, G. Spanakis, A Statutory Article Retrieval Dataset in French, S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 6789–6803. https://doi.org/10.18653/v1/2022.acl-long.468.
URL https://aclanthology.org/2022.acl-long.468/

[17] C. Nguyen, S. T. Luu, T. Tran, A. Trieu, A. Dang, D. Nguyen, H. Nguyen, T. Pham, T. Pham, T.-T. Vo, D.-T. Dol, N.-K. Le, D.-H. Nguyen, N.-C. Le, T.-T. Le, Q. Bui, P. Nguyen, H.-T. Nguyen, V. Tran, L.-M. Nguyen, A Summary of the ALQAC 2023 Competition, 2023 15th International Conference on Knowledge and Systems Engineering (KSE), 2023, pp. 1–6, https://doi.org/10.1109/KSE59128.2023.10299527.

[18] T.-M. Nguyen, H.-T. Nguyen, T.-K. Dao, X.-H. Phan, H. T. Nguyen, T.-H.-Y. Vuong, VLQA: The First Comprehensive, Large, and High-Quality Vietnamese Dataset for Legal Question Answering, ArXiv, Vol. abs/2507.19995, 2025, https://doi.org/10.48550/arXiv.2507.19995.

[19] Z. Xu, S. Escalera, A. Pavão, M. Richard, W.-W. Tu, Q. Yao, H. Zhao, I. Guyon, Codabench: Flexible, Easy-to-use, and Reproducible Meta-benchmark Platform, Patterns, Vol. 3, No. 7, 2022, pp. 100543. https://doi.org/10.1016/j.patter.2022.100543.
URL https://www.sciencedirect.com/science/article/pii/S2666389922001465

[20] S. Robertson, H. Zaragoza, The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends® in Information Retrieval, Vol. 3, No. 4, 2009, pp. 333–389, https://doi.org/10.1561/1500000019.

[21] A. Trotman, A. Puurula, B. Burgess, Improvements to BM25 and Language Models Examined, Proceedings of the 19th Australasian Document Computing Symposium, ADCS '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 58–65, https://doi.org/10.1145/2682862.2682863.

[22] G. M. Rosa, R. C. Rodrigues, R. de Alencar Lotufo, R. Nogueira, Yes, BM25 is a Strong Baseline for Legal Case Retrieval, ArXiv, Vol. abs/2105.05686, 2021, https://doi.org/10.48550/arXiv.2105.05686.

[23] N. Q. Duc, L. H. Son, N. D. Nhan, N. D. N. Minh, L. T. Huong, D. V. Sang, Towards Comprehensive Vietnamese Retrieval-Augmented Generation and Large Language Models, ArXiv, Vol. abs/2403.01616, 2024, https://doi.org/10.48550/arXiv.2403.01616.

[24] N. D. Phuong, ViRanker: A Cross-encoder Model for Vietnamese Text Ranking, 2024.