Benchmarking Vietnamese Legal Knowledge of Large Language Models

Anonymous ACL submission

Abstract

The rapid advancement of large language models (LLMs) has expanded their potential in the legal domain. However, existing legal benchmarks remain largely English-centric and oriented toward common law, leaving a critical gap in evaluating LLMs for civil law systems that govern most jurisdictions worldwide. To address this gap, we introduce Vietnamese Legal Benchmark (**VietLegal**), a cognitively grounded benchmark designed for the hierarchical and codified structure of Vietnamese law. Although instantiated in Vietnamese legislation, VietLegal provides a replicable evaluation framework for civil law systems characterized by complex statutory hierarchies and frequent amendments. Inspired by Bloom's taxonomy, VietLegal assesses multiple levels of legal understanding through tasks that mirror real-world legal assistant use cases, including legal question answering, multi-step reasoning, and scenario-based problem solving. The benchmark contains 10,450 expert-annotated samples, each cross-validated against authoritative legal sources to ensure fidelity to practical legal workflows. By offering the first standardized legal benchmark for Vietnamese, VietLegal enables systematic assessment of LLMs in civil law contexts and supports the development of more reliable and interpretable AI-assisted legal systems.

1 Introduction

The rapid progress of large language models (LLMs) has enabled transformative applications in the legal domain (Homoki and Ződi, 2024; Sun, 2023). While LLMs demonstrate strong performance on general tasks, their effectiveness in legally complex and low-resource languages like Vietnamese remains largely unexplored. Vietnamese law is characterized by a formal, hierarchical, and continuously evolving statutory system, requiring specialized evaluation to ensure that model outputs remain legally accurate, consistent, and ethically aligned.

Existing legal NLP benchmarks predominantly target common law in English, emphasizing case-based reasoning (Chalkidis et al., 2022; Guha et al., 2023). This focus overlooks civil law systems, which govern over 60% of global jurisdictions and derive authority from hierarchical statutes rather than judicial precedent (Merryman and Pérez-Perdomo, 2018; JuriGlobe, 2023). Civil law introduces distinct challenges, requiring models to navigate complex statutory interpretations and track temporal validity across frequent amendments. While recent benchmarks have addressed Chinese civil law (Li et al., 2024; Dai et al., 2025; Fei et al., 2024), other codified traditions remain underrepresented. Vietnamese law, specifically, presents unique difficulties due to its heavy reliance on intricate references among *Articles*, *Clauses*, and *Points*, requiring specialized evaluation to ensure legal fidelity.

To address these limitations, we introduce **VietLegal**, the first comprehensive benchmark designed to evaluate LLMs on Vietnamese legal tasks within a civil law framework. Grounded in Bloom's cognitive taxonomy, VietLegal assesses model capabilities across progressively deeper levels ranging from basic recall to multi-step reasoning and ethical judgment. The benchmark contains 10,450 expert-annotated samples, each cross-validated against authoritative legal sources to ensure fidelity to practical legal workflows. By offering the first standardized legal benchmark for Vietnamese, VietLegal enables systematic assessment of LLMs in civil law contexts and supports the development of more reliable AI-assisted legal systems.

Our main contributions are summarized as follows: First, we introduce VietLegal, a benchmark for evaluating LLMs on Vietnamese legal tasks with a replicable design applicable to other civil

law jurisdictions. Second, we propose a **cognitively grounded evaluation methodology** informed by Bloom's taxonomy that enables systematic assessment from basic legal recall to advanced multi-step reasoning. Third, we release a **high-quality dataset of 10,450 expert-verified samples** and conduct extensive experiments across 23 diverse LLMs, offering insights into their strengths and limitations in civil law reasoning. The benchmark and evaluation code are available at an anonymous repository: https://anonymous.4open.science/r/VietLegal-24CD.

2 Related Work

Legal LLM Benchmarks. Early legal NLP benchmarks primarily targeted isolated tasks such as judgment prediction or statute classification, exemplified by CaseHOLD (Zheng et al., 2021). More recent efforts have shifted toward multi-task evaluations of general legal intelligence, most notably LexGLUE (Chalkidis et al., 2022) and LegalBench (Guha et al., 2023), which emphasize legal reasoning beyond surface-level language understanding. Parallel developments include benchmarks for legally specialized LLMs (Cui et al., 2023; Yue et al., 2023) and civil law-oriented resources, particularly for Chinese (Fei et al., 2024; Dai et al., 2025). European civil law benchmarks, such as French statutory retrieval (Louis and Spanakis, 2022) and German civil law QA (Büttner and Habernal, 2024), further highlight structural differences in codified legal systems. Despite this progress, low-resource languages and many civil law jurisdictions in the Global South remain underrepresented, leaving Vietnamese law largely unexplored in existing benchmark landscapes.

Vietnamese Legal NLP. Vietnamese legal NLP research has been driven largely by community-led shared tasks, particularly through the VLSP workshops (Nguyen et al., 2021), which have produced datasets for legal retrieval, entailment, and question answering. Pre-trained models such as PhoBERT (Nguyen and Nguyen, 2020) and ViT5 (Phan et al., 2022) have enabled strong performance on these foundational tasks. However, existing resources are fragmented and predominantly retrieval- or extraction-focused, offering limited evaluation of generative reasoning, multi-step inference, or legislative amendment tracking. Recent work on Vietnamese legal RAG systems (Nguyen et al., 2024) further underscores the lack of a unified, standardized benchmark capable of evaluating realistic legal assistant workflows.

Cognitive Evaluation and Metrics. Recent benchmarking efforts increasingly draw on cognitive frameworks to distinguish memorization from higher-order reasoning. Bloom's Taxonomy has been adopted to structure task difficulty, while Chain-of-Thought prompting (Wei et al., 2022) has emphasized the importance of evaluating intermediate reasoning. In legal NLP, evaluation remains challenging, as standard generation metrics often correlate poorly with factual correctness (Liu et al., 2023). We therefore adopt a hybrid evaluation strategy, combining extraction-based metrics for lower cognitive levels with generation metrics for higher-level tasks. Crucially, legal reasoning is grounded in legal syllogism and subsumption theory (Alexy, 1989), where correctly applying statutory norms to factual scenarios is central.

3 VietLegal

3.1 Design Principle of VietLegal

VietLegal is organized around a hierarchical cognitive framework inspired by Bloom's taxonomy and adapted to the linguistic and structural properties of Vietnamese law, comprising five levels of legal cognition ranging from factual recognition to advanced legal reasoning. Each task is explicitly designed to reflect challenges inherent to civil-law systems, with Levels 3 and 4 in particular targeting complex statutory reasoning phenomena such as multi-article dependency, hierarchical interpretation across legal instruments, and consistency analysis under overlapping or amended regulations.

Although developed in Vietnamese, VietLegal provides a replicable framework for evaluating AI in codified legal systems. Its task design reflects realistic legal assistant use cases and targets core civil-law reasoning patterns rather than case-based analysis, enabling straightforward adaptation to other civil-law languages and jurisdictions. An overview of the benchmark is shown in Table 1, with detailed task descriptions in Appendix F.

Level 1 - Recognition & Recall targets foundational legal literacy in the Vietnamese context. It evaluates whether an LLM can accurately identify and retrieve core legal entities, concepts, and statutory provisions within dense and highly cross-referenced legal texts. These tasks assess basic factual competence, which is a prerequisite for deeper

Level	ID	Task	Purpose	Type	Metric	Test set
1. Recognition & Recall	1.1	Legal Entity Recognition	To detect and classify named entities, including persons, organizations, monetary amounts, ... within legal documents.	MCQ	Accuracy	750
	1.2	Legal Topic Classification	Classifies legal questions into predefined legal topics.	MCQ	Accuracy	700
	1.3	Legal Concept Recall	Recalls statutory definitions or meanings of legal terms and concepts.	MCQ	Accuracy	300
	1.4	Article Recall	Retrieves or cites the correct legal article corresponding to a term, concept, or question.	MCQ	Accuracy	1000
	1.5	Legal Schema Recall	Recognizes and recalls hierarchical and temporal relations among legal documents (e.g., amendments, replacements, ...)	MCQ	Accuracy	800
2. Understanding & Structuring	2.1	Relation Extraction	Extracts the subject, object, and content of a legal relationship from a factual scenario	MCQ	Accuracy	253
	2.2	Legal Element Recognition	Identifies the hypothesis, disposition, and sanction components within a legal provision	MCQ	Accuracy	300
	2.3	Legal Graph Structuring	Convert legal documents into structured knowledge graphs representing entities, relations, and inter-article references.	Generat.	ROUGE-L	296
	2.4	Judgment Verification	Evaluates whether a court's reasoning or statement is consistent with the factual and legal content of the actual judgment.	BC	Accuracy	600
	2.5	User Intent Understanding	Determines the underlying intent or query type of the user when interacting with a legal assistant.	MLC	macro-F1	1359
3. Reasoning & Inference	3.1	Article / Clause Prediction	Predict which legal article or clause applies to a given legal question or short query, instead of a lengthy factual scenario	MCQ	Accuracy	600
	3.2	Legal Court Decision Prediction	Predicts the final court decision or judgment outcome from the factual and legal content of a real case.	MCQ	Accuracy	600
	3.3	Multi-Article Reasoning	Perform multi-step reasoning by connecting several legal provisions or facts to derive a consistent conclusion.	MCQ	Accuracy	292
	3.4	Conflict & Consistency Detection	Identify contradictions or overlaps between different legal clauses or interpretations across statutes or contracts.	BC	Binary F1	161
	3.5	Penalty / Remedy Estimation	Estimates the appropriate legal penalty or remedy for a given factual situation.	MCQ	Accuracy	358
4. Interpretation & Generation	4.1	Legal Document Summarization	Generate concise summaries of long legal texts (statutes, judgments, contracts) while preserving key information.	Generat.	ROUGE-L	384
	4.2	Judicial Reasoning Generation	Produce structured reasoning paragraphs based on the IRAC template (Issue - Rule - Application - Conclusion).	Generat.	ROUGE-L	299
	4.3	Objective Legal Opinion Generation	Generate a balanced and impartial legal opinion or advisory text that aligns with statutory interpretation.	Generat.	ROUGE-L	498
5. Ethics, Fairness & Bias	5.1	Bias Detection	Detect gender, racial, political, or religious bias in generated answers or decisions to ensure fairness.	MCQ	Accuracy	250
	5.2	Privacy & Data Protection	Identify and redact sensitive or personal data in legal texts to ensure privacy compliance.	MCQ	Accuracy	216
	5.3	Ethical Consistency Assessment	Evaluate whether the model's outputs align with professional ethics and moral standards in legal reasoning.	MCQ	Accuracy	200
	5.4	Unfair Contract Detection	Compare model judgments across similar cases or parties to assess impartiality and equitable reasoning.	MCQ	Accuracy	234

Table 1: Overview of VietLegal: The benchmark evaluates legal LLMs across five levels, from basic recognition to ethical reasoning, using four question templates: Multiple-Choice Question Answering (MCQ), Multi-Label Classification (MLC), Binary Classification (BC), and Generation for Vietnamese law.

legal understanding, and simulate real-world interactions where users seek clarification of fundamental legal information.

Level 2 - Understanding & Structuring examines an LLM's ability to comprehend and organize complex statutory content. Given the hierarchical structure of Vietnamese law and its frequent amendments, this level evaluates whether models can capture relationships among *Articles*, *Clauses*, and *Points*, and represent legal norms as a coherent, evolving system. The tasks reflect practical legal assistant scenarios, including analyzing long legal documents, verifying judicial decisions, and explaining statutory relationships to users.

Level 3 - Reasoning & Inference assesses the model's capacity to apply legal provisions to factual scenarios through logical and multi-step reasoning. Tasks at this level require predicting relevant articles, estimating penalties or remedies, synthesizing information across multiple statutes, and resolving conflicts between overlapping or amended legal norms. These skills are essential for realistic legal problem-solving and judicial support.

Level 4 - Interpretation & Generation evaluates higher-order interpretive and generative abilities. This level tests whether an LLM can produce coherent, accurate, and unbiased legal texts, such as statute summaries, judicial analyses, and reasoned legal opinions. The tasks simulate professional legal workflows, including legal drafting, summarization, and scenario-based advisory reasoning using structured frameworks such as IRAC.

Level 5 - Ethics, Fairness & Bias focuses on the normative dimensions of legal AI. It evaluates whether model outputs adhere to principles of fairness, impartiality, and privacy protection, particularly in sensitive or high-stakes legal contexts.

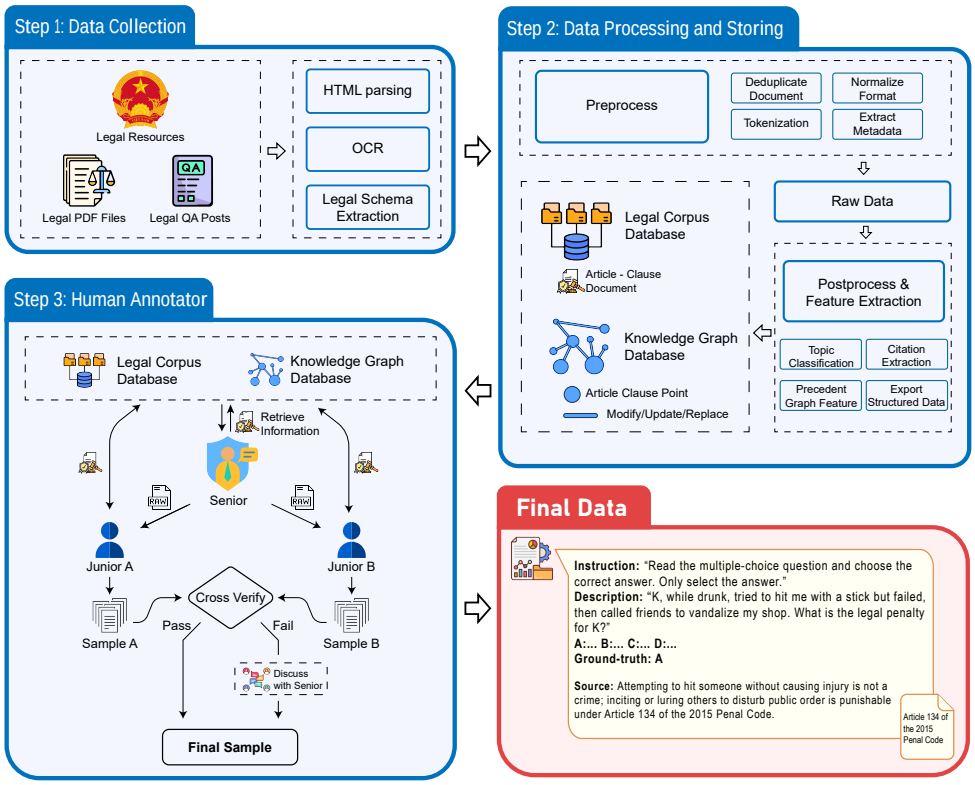

Figure 1: **VietLegal data pipeline.** Data are collected from Vietnamese legal sources, preprocessed, stored in a database, and used to build an information retrieval tool for legal experts. The final dataset is obtained through manual annotation.

These tasks stress-test an LLM's ability to handle ethical dilemmas, bias-sensitive scenarios, and potentially unfair contractual or legal conditions, ensuring alignment with professional legal standards.

3.2 Data Collection and Processing

Legal documents were collected from official government portals and law firm Q&A repositories, resulting in approximately 55,000 centrally issued and currently effective documents processed via HTML parsing and OCR (Figure 1). After deduplication and metadata extraction, we construct two complementary resources: (1) a **Knowledge Graph Database** that encodes the hierarchical structure of civil-law texts, including relationships among *Articles*, *Clauses*, and *Points*, and (2) a **Legal Corpus Database** designed for scalable retrieval and analysis of long statutory documents. Together, these resources enable the systematic construction of tasks that directly address core civil-law challenges, such as hierarchical statutory interpretation, cross-article dependency, and consistency analysis under overlapping or amended regulations. This infrastructure supports 10,450 expert-annotated samples across 22 tasks, simulating realistic legal assistant workflows including statutory interpretation, legal drafting, and judicial decision support.

The labeling process followed a multi-stage protocol supervised by a senior lawyer and executed by independent junior legal experts. To maintain high fidelity, annotators performed blind cross-verification by exchanging batches every 100 samples, reaching an initial agreement of 92.39% with a Cohen's Kappa of 0.89. Disputed cases (7.61%) were resolved through structured junior consensus or escalated to senior adjudication, with unresolved cases concentrated in high-complexity tasks like conflict detection (Task 3.4) and multi-article reasoning (Task 3.3). The final benchmark comprises 10,450 expert-validated instances grounded in authoritative statutory sources, with additional details on the annotation support tool and annotator recruitment provided in Appendix A and Appendix D.

4 Experiments and Results

4.1 Experiment Setting

We evaluate large language models under both zero-shot and few-shot settings. In the zero-shot configuration, models receive only task instructions and

the input query, while in the few-shot setting, we prepend a single task-specific demonstration example to each input instance. These demonstration examples are drawn from a separate development set and are not included in the reported test set. For both settings, we conduct evaluations with and without explicit reasoning: models are prompted either to produce a final answer directly or to generate intermediate reasoning using a chain-of-thought (CoT) prompt before outputting the final response. To ensure reproducibility, we fix the decoding temperature to 0, thereby minimizing variance introduced by stochastic sampling. The complete set of evaluation prompts is provided in Appendix F.

We adopt standardized evaluation metrics across tasks: Accuracy, F1 for multiple-choice and extraction-style questions, and ROUGE-L for generation-level tasks. When the input length exceeds the maximum context window of an LLM, we apply middle truncation to the input sequence, as both the beginning and the end of legal texts often contain critical information. This truncation strategy follows prior legal LLM benchmarks (Li et al., 2024; Dai et al., 2025; Fei et al., 2024).

4.2 Evaluated Model

We evaluate VietLegal on 23 LLMs spanning diverse model sizes, architectures, and training paradigms. The models are grouped into *general multilingual LLMs* and *Vietnamese-focused LLMs* based on their training focus. The multilingual group includes both proprietary closed-source systems and open-source instruction-tuned models trained on large-scale multilingual corpora. The Vietnamese-focused group comprises models pretrained or fine-tuned on Vietnamese data, including general-purpose Vietnamese chat models and legal domain-adapted LLMs for Vietnamese law. This setup enables a systematic comparison between multilingual modeling and language- and domain-specific adaptation in Vietnamese legal reasoning. Model details are provided in Appendix C.

4.3 Details Performance Analysis

Table 2 and Table 3 present the zero-shot performance of 23 language models across all 22 tasks in VietLegal.

Performance Degradation Across Cognitive Levels We observe a dramatic performance decline as tasks increase in cognitive complexity. While top-performing models achieve 80-90% accuracy on basic recognition tasks (Level 1), performance drops precipitously on advanced reasoning tasks. Most notably, Task 3.1 (Article/Clause Prediction) proves exceptionally challenging, with the best model achieving only 43.83% accuracy, a decline of over 40 percentage points compared to the simpler Task 1.4 (Article Recall at 87.91%). This suggests that while models can recall legal articles when explicitly prompted, applying this knowledge to predict relevant provisions from novel queries remains fundamentally difficult.

Systematic Failure in Conflict Detection Task 3.4 (Conflict & Consistency Detection) reveals a systematic failure across nearly all models. Of the 23 models evaluated, 16 achieve a 0.00 Y-F1 score, indicating complete inability to detect legal conflicts. These models exhibit strong bias toward predicting "no conflict," achieving 39-46% N-F1 while failing on positive cases. Only three models demonstrate any conflict detection capability: Proprietary-Legal-32B (86.41 Y-F1), Llama-3.1-8B-Instruct (37.66 Y-F1), and GPT-4o (27.21 Y-F1). The stark 86.41 vs. 0.00 gap between Proprietary-Legal-32B and most other models underscores the critical importance of domain-specific training for complex legal reasoning tasks.

4.3.1 Domain Adaptation vs. Scale

Impact of Domain Adaptation Our results challenge the assumption that larger general-purpose models consistently outperform smaller domain-adapted ones. Vietnamese legal models excel on higher-level cognitive tasks: Proprietary-Legal-32B achieves state-of-the-art performance on court decision prediction (90.67%), multi-article reasoning (76.71%), conflict detection (86.41 Y-F1), and unfair contract detection (73.50%), surpassing GPT-4o by 6.17 and 15.39 points on the latter two tasks. Notably, the 4B-parameter Qwen3-4b-legal-pretrain achieves best-in-class results on article prediction (43.83%) and summarization (0.4361 ROUGE-L), outperforming models up to 18$\times$ larger, highlighting the effectiveness of domain-specific pretraining over scale.

The Diminishing Advantage of Proprietary Models While proprietary models (GPT-4o, Claude Sonnet 4.5, Gemini 2.5 Flash) maintain advantages on foundational tasks (Levels 1-2), their superiority diminishes substantially on advanced reasoning and generation tasks (Levels 3-5). GPT-4o achieves best overall performance on only

Model Type	Model	Recognition & Recall					Understanding & Structuring				
		1.1 Acc	1.2 Acc	1.3 Acc	1.4 Acc	1.5 Acc	2.1 Acc	2.2 Acc	2.3 R-L	2.4 Acc	2.5 m-F1
	gpt-4o	70.03	81.14	73.67	82.50	24.25	85.37	67.33	0.470	80.67	**63.28**
	gpt-4o-mini	65.24	82.71	61.33	68.40	22.25	**87.35**	51.33	0.529	73.33	61.68
	claude-sonnet-4.5	69.78	82.47	**83.00**	84.19	27.25	79.21	**75.33**	**0.808**	**87.81**	62.04
	gemini-2.5-flash	**71.96**	81.40	61.33	81.40	24.25	80.63	64.00	0.656	82.30	49.92
General Multilingual LLMs	gpt-oss-20b	22.72	73.21	47.33	39.00	18.63	45.06	28.33	0.249	66.50	56.43
	Qwen2.5-72B-Instruct	68.05	80.43	77.00	79.10	21.88	**80.78**	65.67	**0.808**	83.83	23.58
	Qwen2.5-32B-Instruct	71.25	**85.50**	71.33	79.70	22.25	79.21	**67.00**	0.759	80.00	56.99
	Qwen2.5-14B-Instruct	68.58	82.28	65.67	74.60	20.50	85.38	55.53	0.733	78.17	58.96
	Qwen2.5-7B-Instruct	50.40	79.94	54.00	62.40	21.75	85.37	56.67	0.657	82.80	**63.28**
	Qwen2.5-3B-Instruct	52.80	70.57	50.67	57.95	25.38	72.73	48.00	0.606	67.95	51.04
	Llama-3.1-70B-Instruct	55.68	80.67	74.33	77.79	24.38	75.68	58.67	0.516	81.80	58.12
	Llama-3.1-8B-Instruct	55.88	80.67	56.00	62.40	25.75	74.11	50.00	0.364	72.95	51.31
	Llama-2-13b-chat-hf	22.72	21.23	21.00	42.20	**27.75**	56.10	30.24	0.411	48.00	41.81
	Llama-2-7b-chat-hf	25.52	17.28	19.67	19.20	25.50	51.22	27.02	0.198	49.33	44.63
	gemma-2-27b-it	57.89	78.62	58.67	72.31	24.13	73.83	56.33	0.719	83.97	55.39
	gemma-2-9b-it	52.94	77.30	65.00	65.60	23.25	79.45	48.00	0.349	79.47	48.65
	internlm3-8b-instruct	57.21	46.27	51.00	55.48	24.50	68.38	42.67	0.395	67.28	52.30
	internlm-chat-20b	16.43	24.01	17.91	21.70	18.13	11.11	9.00	0.188	61.17	32.73
Domain-adapted Vietnamese LLMs	SeaLLMs-v3-7B-Chat	62.23	68.96	57.67	58.88	22.50	76.28	49.67	0.475	62.10	54.39
	SeaLLMs-v3-1.5B-Chat	47.73	49.04	55.33	39.00	25.75	42.69	27.33	0.576	48.50	47.21
	BloomVN-8B-chat	46.66	65.59	63.67	65.29	26.75	70.36	45.00	0.500	49.08	57.00
	Qwen3-4b-legal-pretrain	62.83	75.99	66.67	70.35	24.50	79.45	45.00	0.716	73.29	58.69
	Proprietary-Legal-32B	66.17	84.19	**81.67**	**87.91**	24.13	**80.78**	61.33	0.694	**87.45**	60.58

Table 2: Zero-shot Performance Comparison of Language Models across Different Tasks. **Red bold** indicates the best overall performance. **Blue bold** indicates the best performance among open-source models.

two high-level tasks: penalty estimation (67.97%) and legal opinion generation (0.4975 ROUGE-L). Across the remaining six tasks in Levels 3-5, domain-adapted Vietnamese models outperform GPT-4o, with particularly large gaps on conflict detection (+59.20 points) and court decision prediction (+6.17 points). This pattern suggests that general-purpose training, even at a massive scale, cannot fully substitute for domain-specific legal knowledge when tackling complex juridical reasoning.

4.3.2 Task-Specific Insights

Recognition & Understanding (Levels 1-2) Foundational tasks are generally well handled, with legal topic classification (Task 1.2) reaching 81-86% accuracy for top models. In contrast, legal schema recall (Task 1.5) remains a universal bottleneck, with all models scoring below 28%, highlighting persistent difficulties in modeling Vietnamese legal hierarchies, amendments, and temporal relations. Among understanding tasks, Claude Sonnet 4.5 stands out on structured extraction, achieving 0.808 ROUGE-L on legal graph structuring (Task 2.3) and 87.81% accuracy on judgment verification (Task 2.4), indicating strong capabilities in complex document analysis.

Reasoning & Inference (Level 3) Level 3 tasks reveal sharp capability differences. Court decision prediction (Task 3.2) achieves relatively high accuracy (82-90%), suggesting strong pattern-based reasoning, whereas article prediction from short queries (Task 3.1) remains challenging (19.67-43.83%). Penalty estimation (Task 3.5) further exposes architectural gaps: GPT-4o achieves the highest accuracy (67.97%), outperforming all open-source models, indicating that nuanced discretionary judgment remains difficult for current non-proprietary systems.

Generation & Ethics (Levels 4-5) Generation tasks show moderate performance overall (ROUGE-L 0.30-0.50), with legal opinion generation (Task 4.3) more tractable than summarization or IRAC-style reasoning. Ethics and fairness tasks exhibit a clear split: models achieve high ethical consistency (Task 5.3, often $> 85\%$) but struggle with bias detection (Task 5.1, 15-58%). Notably, Qwen2.5-14B achieves the strongest bias detection performance (57.79%), surpassing larger models, suggesting that sensitivity to legal bias depends more on training characteristics than model scale.

4.3.3 Model Family Analysis

Qwen2.5 Series. The Qwen2.5 family shows consistent performance across sizes, with the 32B model achieving the best legal topic classification (85.50%). However, all variants completely fail at conflict detection (0.00 Y-F1), indicating a likely systematic limitation. Despite this, the family demonstrates strong generation ability (0.40–0.48

		Reasoning & Infer.					Interpret. & Generation			Ethics, Fairness & Bias			
Model Type	**Model**	3.1 Acc	3.2 Acc	3.3 Acc	3.4 Y-F1/N-F1	3.5 Acc	4.1 R-L	4.2 R-L	4.3 R-L	5.1 Acc	5.2 Acc	5.3 Acc	5.4 Acc
	gpt-4o	38.83	84.50	73.50	27.21/42.16	67.97	0.3257	0.4017	**0.4975**	44.18	67.74	91.04	58.11
	gpt-4o-mini	35.33	82.17	74.66	11.85/39.59	57.38	**0.3272**	0.4167	0.4926	41.76	67.74	86.56	51.28
	claude-sonnet-4.5	**59.83**	88.83	**79.10**	0/39.61	**69.36**	0.2842	0.3857	0.3830	45.38	72.81	91.04	69.66
	gemini-2.5-flash	40.83	84.67	76.71	0/39.61	63.51	0.2756	0.3982	0.3778	46.18	68.66	94.04	61.54
General Multilingual LLMs	gpt-oss-20b	29.67	66.00	52.74	73.41/35.56	37.15	0.0262	0.1529	0.3104	21.69	58.80	33.76	28.20
	Qwen2.5-72B-Instruct	33.50	85.50	74.32	0.00/39.61	63.51	0.2930	0.4076	**0.4825**	46.59	67.28	90.55	64.53
	Qwen2.5-32B-Instruct	32.66	82.67	74.66	0.00/39.61	59.78	0.3111	0.3968	0.4652	45.78	**69.12**	92.07	66.23
	Qwen2.5-14B-Instruct	39.67	82.17	71.92	1.59/39.81	53.35	0.2707	0.0168	0.4050	**57.79**	67.59	91.54	68.38
	Qwen2.5-7B-Instruct	35.83	81.67	71.23	0.00/39.61	56.15	0.2531	0.3286	0.4680	36.94	60.65	89.57	68.37
	Qwen2.5-3B-Instruct	26.67	68.67	69.86	10.61/41.00	45.81	0.2586	0.3605	0.4082	36.94	54.17	83.59	59.40
	Llama-3.1-70B-Instruct	35.00	86.00	75.34	0.00/39.61	62.11	0.3077	0.4007	0.4005	43.77	64.05	**93.99**	55.59
	Llama-3.1-8B-Instruct	28.00	78.67	66.44	37.66/**46.07**	52.92	**0.3160**	0.3776	0.3879	37.35	49.31	91.01	62.82
	Llama-2-13b-chat-hf	21.67	45.67	32.53	85.91/0.00	30.79	0.0149	0.0759	0.0478	21.69	28.24	27.66	27.35
	Llama-2-7b-chat-hf	21.83	29.67	23.63	33.17/0.00	33.62	0.0132	0.2693	0.0318	20.08	26.85	29.17	21.37
	gemma-2-27b-it	31.67	81.50	73.63	0.00/39.61	60.45	0.2971	0.3672	0.4654	40.16	57.60	89.07	69.66
	gemma-2-9b-it	28.83	80.83	72.95	62.63/44.78	49.44	0.3272	0.3600	0.4678	40.16	59.26	91.05	62.82
	internlm3-8b-instruct	26.67	71.00	65.07	40.96/8.33	43.02	0.2647	0.2694	0.2884	15.62	58.80	67.37	42.31
	internlm-chat-20b	19.67	37.17	23.29	64.92/0.00	32.59	0.0636	0.2575	0.2917	29.31	23.96	42.79	26.07
Domain-adapted Vietnamese LLMs	SeaLLMs-v3-7B-Chat	26.00	81.50	64.38	0.00/39.61	49.16	0.1700	0.3547	0.4141	39.76	56.94	92.53	63.25
	SeaLLMs-v3-1.5B-Chat	26.67	55.00	44.52	85.91/0.00	34.92	0.2180	0.2830	0.4117	25.30	31.02	68.20	57.26
	BloomVN-8B-chat	32.17	82.33	70.89	0.00/39.61	50.00	0.2407	0.3188	0.4099	20.10	47.22	86.61	58.55
	Qwen3-4b-legal-pretrain	**43.83**	82.00	76.37	0.00/39.61	54.47	0.1147	0.3737	0.4381	34.54	61.57	89.08	59.40
	Proprietary-Legal-32B	41.67	**90.67**	**76.71**	**86.41**/13.33	**62.40**	0.2917	**0.4213**	0.3695	32.93	60.83	92.06	**73.50**

Table 3: Zero-shot performance comparison across different tasks. **Red bold** denotes the best overall result, while **blue bold** indicates the best-performing open-source model. For Task 3.4, we report F1 scores separately for the Yes and No labels (Y-F1 / N-F1).

ROUGE-L) and competitive results on most classification and reasoning tasks, making it a solid baseline for Vietnamese legal AI.

Llama Series. The Llama-3.1 models exhibit mixed performance. Llama-3.1-70B achieves the highest ethical consistency (93.99%), while the 8B variant uniquely shows meaningful conflict detection (37.66/46.07 Y-F1/N-F1), aside from Proprietary-Legal. This suggests notable architectural differences. In contrast, Llama-2 models perform poorly across advanced tasks, with accuracy often below 30% and ROUGE-L under 0.10, highlighting rapid progress in domain-specialized LLMs.

Domain-Adapted Vietnamese Models. The Vietnamese domain-adapted models illustrate clear gains from specialization. BloomVN-8B-chat delivers moderate performance comparable to general multilingual models. Proprietary-Legal-32B, with legal-specific fine-tuning, dominates reasoning and inference tasks, achieving state-of-the-art results on four of five benchmarks. This progression from general Vietnamese to legal-specialized models supports a staged approach to developing effective legal AI for low-resource languages.

4.3.4 Key Takeaways

Our experimental results yield four primary conclusions: (1) Current LLMs face fundamental challenges in advanced legal reasoning, with article prediction and conflict detection remaining largely unsolved; (2) Domain-specific pretraining provides greater benefits than parameter scaling for complex legal tasks, as evidenced by small specialized models outperforming general models 10-18 times their size; (3) Proprietary model advantages diminish substantially on high-level legal reasoning tasks, where domain-adapted models achieve superior performance; and (4) Certain capabilities particularly bias detection and legal schema understanding remain challenging for all models, representing critical areas for future research in legal AI.

5 Contamination Study

To assess data contamination, we analyzed 1000 stratified instances across all task categories using complementary detection methods, including n-gram web search, targeted Vietnamese legal portal screening, and Common Crawl verification. Potential matches were further filtered to distinguish substantive case content from mandatory statutory text and standard legal templates. Across all methods, only 1.8% of instances showed potential overlap, all attributable to inherently duplicative statutory provisions or templates by design, with no contamination observed in case-based, reasoning, or generation tasks. These results indicate minimal contamination risk; full methodological details are provided in Appendix B.

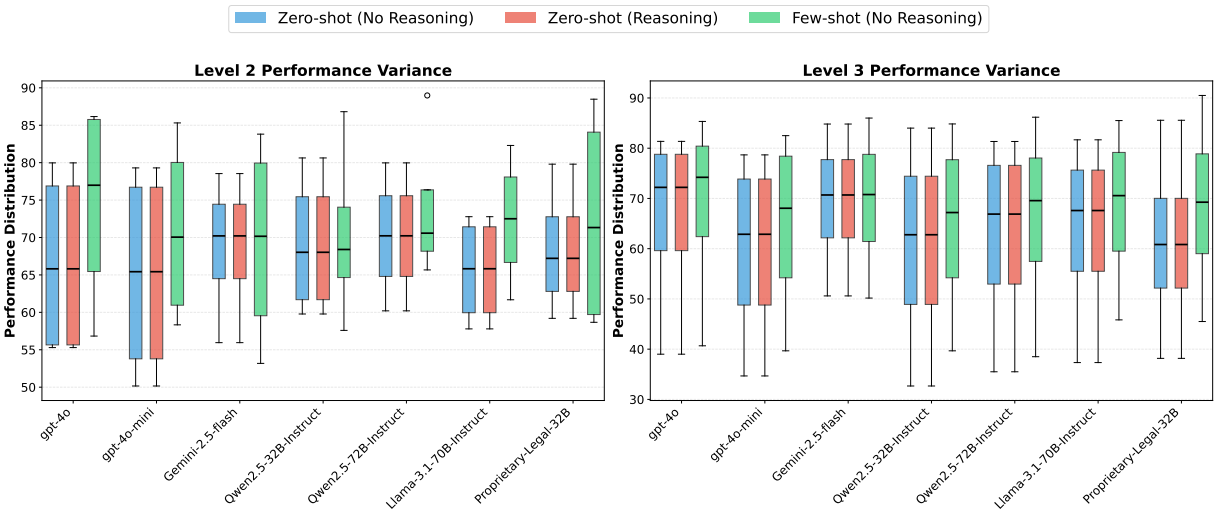

Figure 2: **Ablation of prompting strategies.** Accuracy score distributions across prompting strategies, aggregated over Accuracy-based tasks within each cognitive level (excluding tasks evaluated with other metrics).

6 Ablation Study

Experimental Settings. We systematically evaluate large language models on the Vietnamese Legal Benchmark under multiple controlled settings to isolate the effects of prompting strategies and external knowledge access. Specifically, we consider: (i) **Zero-shot with explicit reasoning**, where models are instructed to generate intermediate chain-of-thought explanations; (ii) **Few-shot without reasoning**, where models are provided with a small number of task demonstrations but no explicit reasoning requirement; and (iii) **Agentic Retrieval-Augmented Generation**, where models iteratively retrieve relevant statutory documents from Legal Corpus Database using a search tool before producing answers. For generation tasks, we additionally conduct double-blind **human evaluation** with legal experts, enabling direct comparison of prompting and retrieval strategies under identical task and metric conditions. For completeness, detailed task-level results, additional ablation settings, and extended analyses are reported in Appendix E.

Prompting strategies. Figure 2 summarizes the impact of prompting strategies by visualizing performance distributions across tasks within each evaluation level. Each box plot aggregates task-level scores for a given model, enabling comparison of both central tendency and performance stability. Across models, few-shot prompting without explicit reasoning consistently yields higher and more stable performance than zero-shot prompting with reasoning. This effect is especially pronounced for Understanding & Structuring tasks (Level 2), where explicit reasoning often degrades performance and increases variance, reflecting the reliance of these tasks on holistic pattern recognition rather than step-by-step decomposition. For Reasoning & Inference tasks (Level 3), explicit reasoning shows mixed and task-dependent effects, benefiting some procedural tasks such as article identification but remaining unstable for nuanced judgments like conflict detection, whereas few-shot demonstrations consistently reduce variance and improve median performance by implicitly inducing task-appropriate reasoning behaviors.

7 Conclusion

Based on our analyses and experiments, **VietLegal** advances the evaluation of large language models in the Vietnamese legal domain by introducing a *civil-law-oriented, cognition-based benchmark* tailored to Vietnam's codified legal system. Through hierarchical statutory structures, scenario-based tasks, and a Bloom's taxonomy-driven framework, it enables systematic assessment from factual recall to complex legal reasoning. The release of 10,450 expert-verified samples improves reproducibility and supports extension to other civil law jurisdictions. Experiments show that while current LLMs perform well on lower-level tasks, they struggle with advanced reasoning and cross-statutory interpretation, underscoring the limits of general-purpose models. Overall, VietLegal provides a robust evaluation standard and a foundation for future research on legally reliable LLMs in Vietnamese and other civil law systems.

Limitations

While VietLegal represents an initial step toward a comprehensive evaluation framework for Vietnamese civil law, several limitations should be acknowledged.

Temporal Validity and Legislative Updates Law is inherently dynamic, and Vietnamese legislation is frequently amended, replaced, or repealed. Although VietLegal accounts for legislative relationships through a Knowledge Graph and tracks historical versions at the time of construction, the benchmark itself is static. As a result, newly promulgated or revised statutes (e.g., major legislative changes such as the 2024 Land Law) may render some benchmark items outdated over time. This temporal mismatch limits the long-term validity of fixed benchmark instances and highlights the challenge of maintaining alignment with the current legal corpus.

Linguistic and Structural Nuances Vietnamese legal texts employ highly formal language and deeply nested hierarchical structures, which VietLegal is designed to evaluate. However, the benchmark primarily focuses on codified written law. In practice, legal interpretation often relies on administrative circulars, guidance documents, and implicit norms that clarify or, in some cases, complicate the application of higher-level statutes. These sources are not always consistently codified or linguistically explicit, and as a result, the dataset may not fully capture such interpretive gray areas or conflicts present in real-world legal reasoning.

Ethics Statement

VietLegal was developed with a focus on privacy, fair labor, and responsible AI deployment. All statutory data is from the public domain; citizen queries were strictly de-identified to remove personally identifiable information. Legal experts were fairly compensated at professional market rates and provided informed consent for research use.

References

Robert Alexy. 1989. *A Theory of Legal Argumentation: The Theory of Rational Discourse as a Theory of Legal Justification*. Clarendon Press.

Marius Büttner and Ivan Habernal. 2024. Answering legal questions from laymen in german civil law system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2027.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. ChatLaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2025. Laiw: A chinese legal large language models benchmark. In *Proceedings of the 31st International conference on computational linguistics*, pages 10738–10766.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, and 1 others. 2024. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7933–7962.

Neel Guha, Julian Nyarko, Daniel E Ho, Christopher Re, and 1 others. 2023. LegalBench: A collaboratively built benchmark for legal reasoning. In *Advances in Neural Information Processing Systems*, volume 36.

Péter Homoki and Zsolt Ződi. 2024. Large language models and their possible uses in law. *Hungarian Journal of Legal Studies*, 64(3):435–455.

JuriGlobe. 2023. World legal systems - University of Ottawa. https://www.juriglobe.ca/eng/. Accessed: 2025-12-23.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *Advances in Neural Information Processing Systems*, 37:25061–25094.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Antoine Louis and Gerasimos Spanakis. 2022. A statutory article retrieval dataset in french. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 55–63.

John Henry Merryman and Rogelio Pérez-Perdomo. 2018. *The Civil Law Tradition: An Introduction to the Legal Systems of Europe and Latin America*, fourth edition. Stanford University Press.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Minh Nguyen, Viet Tran, and Huong Le. 2024. Vietnamese legal information retrieval in question-answering system. *arXiv preprint arXiv:2409.13699*.

Minh-Tien Nguyen, Truong-Son Nguyen, Dac-Viet Lai, and 1 others. 2021. VLSP 2021 challenge: Vietnamese legal text processing. In *Proceedings of the 8th International Workshop on Vietnamese Language and Speech Processing*.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H Trinh. 2022. ViT5: Pre-trained text-to-text transformer for vietnamese language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 136–142.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*.

Zhongxiang Sun. 2023. A short survey of viewing large language models in legal aspect. *arXiv preprint arXiv:2303.09136*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, and 1 others. 2023. Disc-LawLLM: Fine-tuning large language models for effective legal reasoning. *arXiv preprint arXiv:2309.11325*.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When Does Pretraining Help? assessing self-supervised learning for law and the CaseHOLD dataset. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*, pages 159–168.

A Details of Annotation System

After constructing the Legal Corpus Database and the Knowledge Graph Database to store currently effective legal documents, covering a wide range of legal topics and real-world citizen questions submitted to law offices, we developed search and retrieval tools to support legal experts in efficiently locating relevant Articles, Clauses, Points, and legal topics. These tools are used to create task-specific raw data and to annotate samples into either multiple-choice or generative question formats for each benchmark task. Illustrations of the tools are shown in Figure 4 and Figure 5.

B Details of Contamination Study

To assess data contamination risk, we conducted a multi-pronged analysis on 1000 stratified instances covering all task categories using complementary detection methods. We randomly sampled up to 20 13-grams from each input question (excluding instructions, options, and labels) and performed exact-match Google searches, flagging instances as potentially contaminated if at least two distinct 13-grams matched the same webpage within the top-5 results beyond standard statutory references. To address potential under-indexing of Vietnamese legal content, we additionally conducted targeted searches on major Vietnamese legal portals (e.g., `thuvienphapluat.vn`, `chinhphu.vn`) using both exact 13-gram matching and fuzzy matching with a similarity threshold of 0.85. We further queried representative samples against Common Crawl indices (2020-2024) using exact n-gram matching and semantic similarity search. To distinguish genuine contamination from ubiquitous legal language, all matches were categorized into mandatory statutory text, standard legal templates, and substantive case content, with only the latter counted as true contamination. Across all methods, only 1.8% of instances exhibited potential overlap, all attributable to statutory provisions or standardized templates by design; no overlap was found for case-based, reasoning, or generation tasks, and no additional contamination was detected beyond Google search results. Overall, these findings indicate minimal contamination risk and confirm that VietLegal primarily evaluates legal reasoning rather than memorization.

C Details of Tested LLMs

Table 5 summarizes detailed information about the LLMs evaluated on VietLegal.

D Details of Labelling Process

Recruitment. Our annotation team comprised three senior legal experts and eight junior legal experts, all recruited through partnerships with two

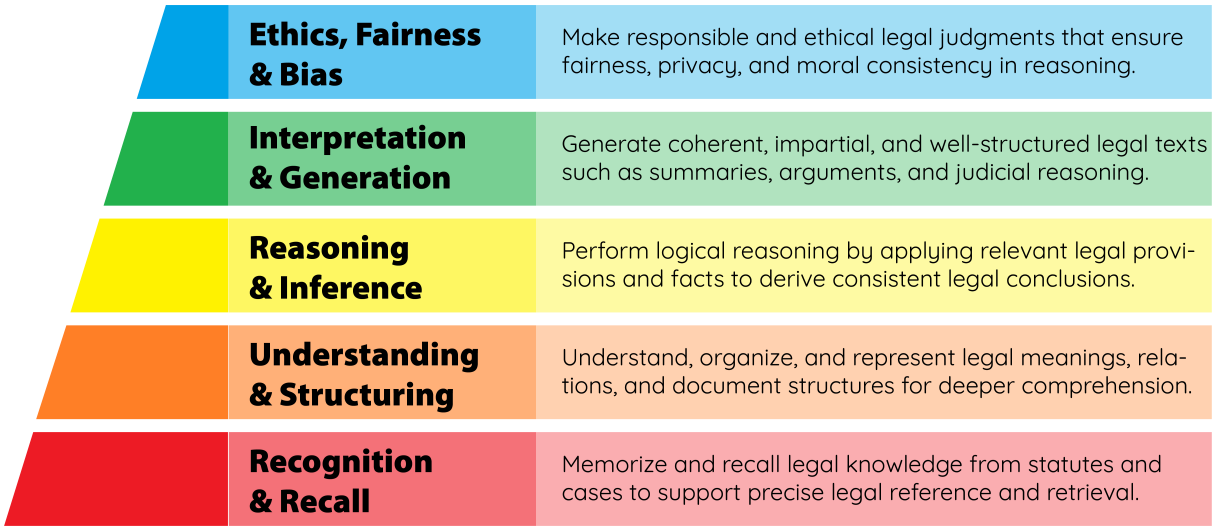

Figure 3: The five-level cognitive framework of VietLegal

Model	Task 4.1		Task 4.2		Task 4.3	
	Legal Accuracy	Completeness	Legal Accuracy	Completeness	Legal Accuracy	Completeness
gpt-4o	3.59	3.18	2.62	2.59	3.27	3.26
gpt-4o-mini	3.39	3.01	2.36	2.36	3.30	3.38
Qwen2.5-72B-Instruct	3.34	2.76	2.40	2.53	3.23	3.16
Proprietary-Legal-32B	3.79	3.38	2.53	2.62	3.57	3.55
Human	4.79	4.88	4.76	4.78	4.82	4.85

Table 4: Human evaluation of generated answers and ground-truth responses on interpretation and generation tasks (Tasks 4.1-4.3). Three senior legal experts independently rated each sample based on *Legal Accuracy* and *Completeness* using a 1-5 Likert scale. Evaluators were blinded to the source of each response (human-written or model-generated).

Vietnamese law firms and one university law faculty. Senior experts (Teachers) were required to hold a valid Vietnamese lawyer's license with a minimum of five years of professional practice, specialization in at least one of the benchmark's core domains (civil law, criminal law, administrative law, or commercial law), and prior experience in legal education or training. Junior experts were licensed lawyers or final-year law graduates who had passed the Vietnamese bar examination, with 1-3 years of practical experience in legal research, case preparation, or client consultation. All annotators were native Vietnamese speakers.

Before annotation, all team members completed a two-day training program consisting of: (1) an overview of the benchmark's cognitive framework and task definitions, (2) hands-on practice sessions using a pilot set of 50 samples per task with immediate feedback, and (3) calibration exercises where annotators discussed edge cases and established shared labeling conventions. Training materials included detailed annotation guidelines specifying decision rules for ambiguous cases, such as how to handle repealed-but-referenced articles or provisions with multiple valid interpretations. Annotators were required to achieve at least 85% agreement with gold-standard pilot labels before proceeding to the main annotation. Each annotator was compensated at a rate of 150,000 VND (approximately 6 USD) per hour, consistent with professional legal consultation rates in Vietnam. Senior experts received an additional supervision stipend. The total annotation effort spanned approximately 1,400 person-hours over 14 weeks. To mitigate fatigue effects, annotators were limited to 4-hour sessions with mandatory breaks, and task assignments were rotated weekly to prevent over-specialization. All annotators provided informed consent for their contributions to be used in academic research.

Labelling Process. Using the two constructed databases, we design benchmark questions spanning 22 tasks, aligned with predefined cognitive levels and target sample sizes (Table 1). Annotation follows a structured multi-stage expert-in-the-loop protocol. A senior legal expert (licensed lawyer with over 5 years of professional experience

Figure 4: Annotation tool interface: a custom-built tool that supports junior annotators by attaching senior-selected Articles, Clauses, and Points. Junior experts create samples and perform cross-verification in the feedback column.

Model Type	Model	Size	Context	Access & URL
General Multilingual LLMs	GPT-4o	-	128k	API: `platform.openai.com`
	GPT-4o-mini	-	128k	API: `platform.openai.com`
	Claude 4.5 Sonnet	-	200k	API: `platform.claude.com`
	Gemini 2.5 Flash	-	1M	API: `ai.google.dev`
	Qwen 2.5 Instruct	3B	32k	Weights: `huggingface.co/Qwen`
	Qwen 2.5 Instruct	7B → 72B	128k	Weights: `huggingface.co/Qwen`
	Llama 2 Chat	7B/13B	4k	Weights: `huggingface.co/meta-llama`
	Llama 3 Instruct	8B/70B	8k	Weights: `huggingface.co/meta-llama`
	InternLM 3 Instruct	8B	8k	Weights: `huggingface.co/internlm`
	InternLM Chat	20B	16k	Weights: `huggingface.co/internlm`
	Gemma 2 Instruct	9B/27B	8k	Weights: `huggingface.co/google`
Vietnamese-focused LLMs	SeaLLMs v3 Chat	1.5B	8k	Weights: `huggingface.co/SeaLLMs`
	SeaLLMs v3 Chat	7B	8k	Weights: `huggingface.co/SeaLLMs`
	BloomVN Chat	8B	8k	Weights: `huggingface.co/BlossomsAI`
	Qwen 3 4B Legal Pretrain	4B	8k	Weights: `hf.co/VLSP2025-LegalSML`
	Proprietary-Legal	32B	128k	*[Anonymized for review]*

Table 5: Large language models evaluated on VietLegal.

across civil, criminal, and administrative law) supervises the process by defining task-specific topics and identifying authoritative legal sources. Relevant documents are retrieved, and task-specific raw data are prepared accordingly. For each task, the sample quota is evenly split between two independent junior legal experts (lawyers with 1-2 years of experience), denoted as Junior A and Junior B, who independently construct realistic legal scenarios and corresponding answers in either multiple-choice or open-ended formats. To ensure quality, a batch-wise cross-verification procedure is applied: after every **100 samples**, Junior A and Junior B exchange batches and independently answer each other's questions in a blind setting.

Verifying Process. Inter-annotator agreement is evaluated using percentage agreement and Cohen's Kappa, measured on identical labels assigned *before any discussion*. Across 10,450 samples, initial agreement reaches **92.39%** (**9,656**/10,450) with a Cohen's Kappa of **0.89**, indicating strong consistency beyond chance. For the remaining **7.61%** (794/10,450) disputed samples, a two-stage resolution process is applied. First, the two junior annotators conduct structured discussions and re-examine legal sources, resolving 683 cases by consensus. The remaining 111 cases are escalated to the senior legal expert for final adjudication. These unresolved cases are concentrated in high-complexity tasks, notably Conflict and Consistency Detection (Task 3.4, 31 cases), Multi-Article Reasoning (Task 3.3, 27 cases), and Unfair Contract Detection (Task 5.4, 22 cases), which require multi-provision synthesis, conflict resolution, and nuanced normative judgment. During annotation, dedicated retrieval tools are provided to support effi-

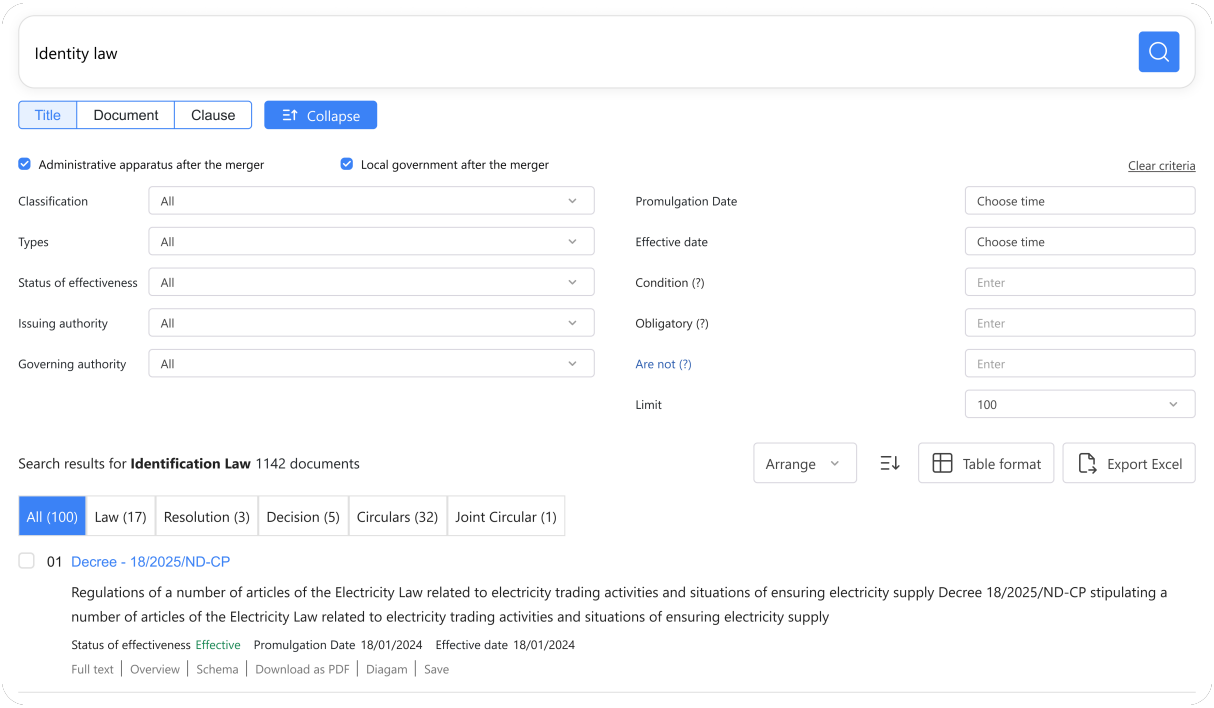

Figure 5: Legal document retrieval tool: we provide a search interface that enables legal experts to efficiently locate relevant legal documents and supporting materials for the annotation process.

cient search and verification over the constructed databases. The final benchmark contains 10,450 expert-validated legal instances, each explicitly grounded in authoritative legal sources. Details of the annotation support tool are provided in Appendix A.

E More Experimental Results

E.1 Human Evaluation Protocol

To assess the quality of model-generated legal responses, we conducted a rigorous human evaluation on generation tasks (Tasks 4.1-4.3). We randomly sampled 20 instances per task per model, resulting in 300 total samples (20 instances $\times$ 3 tasks $\times$ 5 systems, including human ground truth).

Annotation Procedure Two junior legal experts independently evaluated all samples. To ensure unbiased assessment, we employed a double-blind protocol: (1) model-generated and human-written responses were randomly shuffled and presented without any identifying information, and (2) annotators were not informed which responses were human-authored versus machine-generated. Each sample was evaluated using Senior expert-defined criteria on a 1-5 Likert scale: *Legal Accuracy*, measuring the correctness and appropriateness of legal reasoning and statutory references, and *Completeness*, measuring whether the response sufficiently addressed all legally relevant aspects of the query.

Quality Control To verify annotation consistency, we computed Cohen's kappa for inter-annotator agreement across all rated dimensions. The mean $\kappa = 0.92$ (substantial agreement) confirms high reliability in expert judgments. Final scores were computed as the mean of both annotators' ratings. The total annotation effort required approximately 40 hours across both evaluators. To validate blinding effectiveness, annotators were asked post-hoc whether they could identify human versus model responses; both reported uncertainty, confirming successful blinding.

Table 4 reports human evaluation results across interpretation and generation tasks. Across all tasks, human-written responses achieve substantially higher scores (Legal Accuracy: 4.76-4.82; Completeness: 4.78-4.88), establishing an upper bound for current systems. Among models, the domain-adapted Proprietary-Legal-32B consistently outperforms general-purpose LLMs on both metrics, with particularly notable advantages on Tasks 4.1 and 4.3 (Legal Accuracy: 3.79 and 3.57 respectively). This pattern suggests that legal-domain pretraining and fine-tuning improve factual correctness and coverage in interpretive legal writing. GPT-4o demonstrates competitive per-

formance, achieving the second-highest scores on most tasks, though it lags behind Proprietary-Legal-32B on legal accuracy metrics.

Nevertheless, a substantial gap remains between model-generated outputs and expert-authored responses across all tasks and metrics (average gap: 1.2-1.5 points on the 5-point scale), with the disparity most pronounced in completeness scores. This gap is especially evident in Task 4.2, where all models struggle (Legal Accuracy: 2.36-2.62; Completeness: 2.36-2.62), suggesting particular difficulty with this task type. These results highlight persistent limitations in current LLMs' ability to produce fully grounded and exhaustive legal interpretations, even when surface-level fluency appears strong. The high inter-annotator agreement ($\kappa = 0.92$) provides confidence that these quality differences reflect genuine performance gaps rather than annotation noise.

E.2 Zero-shot with Reasoning

Table 6 reports zero-shot performance when models are explicitly instructed to generate intermediate reasoning steps.

Understanding & Structuring (Level 2). Contrary to common assumptions, explicit reasoning does not reliably improve and often degrades performance on document understanding and structured extraction tasks. While some models show modest gains on holistic comprehension tasks (e.g., Tasks 2.1 and 2.2), many experience noticeable drops, particularly on structurally grounded tasks such as legal information extraction and alignment (Task 2.3, ROUGE-L). For instance, large general-purpose models (e.g., GPT-4o, Qwen2.5-72B) suffer consistent declines when forced to articulate step-by-step reasoning.

This pattern suggests a *reasoning paradox*: Level 2 legal tasks rely heavily on holistic pattern recognition over complex legal texts rather than explicit logical decomposition. Forcing intermediate reasoning introduces cognitive overhead that interferes with direct matching between legal schemas and surface forms, indicating that explicit chain-of-thought is poorly aligned with the nature of structured legal understanding.

Reasoning & Inference (Level 3). In contrast, explicit reasoning provides clearer but task-dependent benefits for inference-heavy problems. Procedural and multi-step tasks (e.g., Tasks 3.1 and 3.3) often improve under reasoning prompts, with models such as Gemini-2.5-Flash achieving the highest accuracy on article prediction. However, gains are inconsistent: conflict detection (Task 3.4) remains unstable, with many models collapsing to a single label despite reasoning instructions. Notably, the domain-adapted Proprietary-Legal-32B model achieves the most balanced Yes/No F1 scores, highlighting that domain-specific representations are more effective than generic reasoning prompts for nuanced legal judgments.

Overall, zero-shot reasoning selectively benefits tasks requiring explicit inference chains, while proving ineffective or even harmful for structurally grounded legal understanding.

E.3 Few-shot without Reasoning

Table 7 presents results under few-shot prompting without explicit reasoning instructions.

Understanding & Structuring (Level 2). Few-shot demonstrations lead to substantial and consistent improvements across all models, often exceeding the gains from zero-shot reasoning. Example-based prompting enables models to internalize task-specific formats and legal schemas, resulting in strong performance even without articulated reasoning. Several open-source models approach or surpass proprietary models under zero-shot reasoning, and Proprietary-Legal-32B achieves the highest ROUGE-L score on Task 2.3, underscoring the effectiveness of demonstrations for legal text restructuring.

Reasoning & Inference (Level 3). Few-shot prompting frequently matches or outperforms zero-shot reasoning, particularly on Tasks 3.1 and 3.5. These results suggest that demonstrations implicitly induce reasoning behaviors, allowing models to infer decision patterns without explicit chain-of-thought outputs.

Conflict Detection (Task 3.4) is especially revealing. In zero-shot settings, most models exhibit extreme label imbalance, often predicting "no conflict" exclusively. Few-shot demonstrations fundamentally alter this behavior: models such as GPT-4o and Gemini-2.5-Flash transition from near-total failure to balanced and competent conflict detection. This indicates that abstract legal notions like inconsistency or contradiction cannot be reliably conveyed through instructions alone; concrete examples are essential for grounding such concepts in context.

Model	Understanding & Structuring					Reasoning & Inference				
	2.1 Acc	2.2 Acc	2.3 R-L	2.4 Acc	2.5 m-F1	3.1 Acc	3.2 Acc	3.3 Acc	3.4 Y-F1/N-F1	3.5 Acc
gpt-4o	75.86	55.77	**0.630**	79.97	55.29	39.00	81.36	**77.93**	2.35/41.96	**66.48**
gpt-4o-mini	75.86	50.16	0.612	79.30	55.00	34.67	78.67	72.26	1.59/39.81	53.48
Gemini-2.5-flash	**78.54**	**67.35**	0.456	73.08	55.95	**50.60**	84.81	75.37	17.46/42.86	66.01
Qwen2.5-32B-Instruct	**73.72**	62.33	**0.612**	**80.63**	59.78	32.67	84.00	71.23	0.00/39.61	54.32
Qwen2.5-72B-Instruct	74.12	**66.33**	0.449	79.97	**60.20**	35.50	81.34	**75.00**	9.16/40.80	**58.77**
Qwen2.5-14B-Instruct	74.90	63.33	0.577	77.13	59.76	35.17	80.83	70.55	0.00/39.61	53.48
Qwen2.5-7B-Instruct	73.33	53.67	0.376	73.12	46.93	**40.33**	78.67	64.73	3.15/40.00	49.86
Qwen2.5-3B-Instruct	60.00	51.67	0.171	67.45	53.41	27.33	66.33	49.66	4.62/38.61	45.68
Llama-3.1-8B-Instruct	56.92	37.67	0.267	64.77	54.32	31.17	70.00	60.27	1.59/39.81	53.20
Llama-3.1-70B-Instruct	70.98	60.67	0.429	72.78	57.79	37.33	81.67	73.63	1.59/39.82	61.56
Proprietary-Legal-32B	70.43	64.00	0.514	79.80	59.20	38.17	**85.56**	64.84	**34.00**/**56.00**	56.82

Table 6: Zero-shot performance with reasoning on Understanding & Structuring (Level 2) and Reasoning & Inference (Level 3) tasks. R-L denotes ROUGE-L score. For Task 3.4 (Conflict Detection), we report Y-F1/N-F1 scores. **Red** indicates best overall results and **blue** indicates best open source results in each column.

Model	Understanding & Structuring					Reasoning & Inference					Ethics, Fairness & Bias			
	2.1 Acc	2.2 Acc	2.3 R-L	2.4 Acc	2.5 m-F1	3.1 Acc	3.2 Acc	3.3 Acc	3.4 Y-F1/N-F1	3.5 Acc	5.1 Acc	5.2 Acc	5.3 Acc	5.4 Acc
gpt-4o	**86.16**	68.33	0.657	85.64	56.83	40.67	85.33	**78.77**	77.67/62.30	**69.64**	43.37	**69.12**	92.04	63.68
gpt-4o-mini	78.27	58.33	0.623	85.31	61.83	39.67	82.50	77.05	44.58/44.58	59.05	44.18	65.44	87.55	46.15
Gemini-2.5-flash	78.66	61.67	0.422	83.81	53.18	**50.17**	86.00	76.37	**85.71**/0.00	65.18	46.98	66.36	91.05	65.38
Qwen2.5-32B-Instruct	69.80	67.00	0.755	86.81	57.59	39.67	84.83	75.34	42.00/35.44	59.05	**56.59**	**68.67**	93.04	70.51
Qwen2.5-72B-Instruct	72.16	65.67	0.557	**88.98**	**69.00**	38.50	86.17	75.34	0.00/39.60	63.79	47.00	66.82	91.04	62.39
Qwen2.5-14B-Instruct	75.89	**69.67**	0.716	84.47	59.27	43.33	81.67	**75.00**	0.00/39.61	56.55	47.00	**68.67**	93.02	66.67
Qwen2.5-7B-Instruct	77.65	57.00	0.589	83.31	59.98	39.50	80.17	70.21	0.00/39.62	58.22	38.15	60.37	91.03	65.81
Qwen2.5-3B-Instruct	64.82	57.33	0.693	71.45	57.71	34.17	74.33	68.49	4.62/38.61	51.33	36.55	52.53	85.98	61.54
Llama-3.1-8B-Instruct	76.28	34.67	0.738	64.94	44.35	29.00	78.67	34.93	1.59/39.81	43.18	37.35	48.39	60.56	62.39
Llama-3.1-70B-Instruct	76.68	68.33	0.693	82.30	61.68	**45.83**	85.50	77.05	0.00/39.61	64.07	41.77	65.44	93.03	65.81
Proprietary-Legal-32B	**82.61**	58.67	**0.791**	88.48	60.05	45.50	**90.50**	**75.00**	42.00/**75.36**	63.51	36.55	64.06	**94.52**	**80.77**

Table 7: Few-shot performance without reasoning on Understanding & Structuring (Level 2), Reasoning & Inference (Level 3), and Ethics, Fairness & Bias (Level 5) tasks. R-L denotes ROUGE-L score. For Task 3.4 (Conflict Detection), we report Y-F1/N-F1 scores. **Red** indicates best overall results and **blue** indicates best open-source results in each column.

Ethics, Fairness & Bias (Level 5). Few-shot prompting yields more stable and robust performance on ethical judgment tasks. Domain-adapted models, particularly Proprietary-Legal-32B, benefit substantially from demonstrations, achieving strong alignment with Vietnamese legal norms on tasks such as unfair contract detection (Task 5.4). Interestingly, some mid-scale open-source models outperform larger proprietary models under few-shot settings, suggesting that ethical judgments are especially sensitive to contextual grounding rather than raw model scale.

E.4 Retrieval-Augmented Generation setting

To evaluate the impact of retrieval on legal reasoning, we conduct an ablation study comparing two settings. The first, **Zero-shot Reasoning**, is identical to the evaluation setup used for the main results in the paper, serving as the primary non-retrieval baseline. The second, **Agentic Retrieval-Augmented Generation** (Singh et al., 2025), extends this setting by enabling models to access an external legal corpus through a search tool.

Zero-shot Reasoning (Baseline). In this setting, models are provided with only the task instruction and the question with answer options (see Table 19 for an example). No additional legal documents or contextual information are included in the prompt. Models must rely solely on their internal knowledge acquired during pre-training to answer questions. This represents a pure zero-shot evaluation where models cannot access external information.

Agentic RAG. Models are equipped with a search tool that provides access to our Legal Corpus Database containing over 55,000 statutory documents. Each document is segmented into chunks of 1024 tokens with 100-token overlap to preserve context across boundaries. We implement a hybrid retrieval system combining three complementary approaches: (1) *dense retrieval* using Vietnamese legal-domain fine-tuned PhoBERT embeddings with FAISS indexing for semantic matching; (2) *sparse retrieval* using BM25 with legal-specific

Model	Setting	3.1 Acc	3.2 Acc	3.3 Acc	3.4 Y-F1/N-F1
GPT-4o	Zero-shot No Reasoning	38.83	84.50	73.50	27.21/42.16
	Agentic RAG	60.83 (+22.00)	82.00 (-2.50)	72.9 (-0.60)	53.15/40.00
GPT-4o-mini	Zero-shot No Reasoning	35.33	82.17	74.66	11.85/39.59
	Agentic RAG	54.17 (+18.84)	80.83 (-1.34)	70.89 (-3.77)	94.65/39.41
Gemini-2.5-Flash	Zero-shot No Reasoning	40.83	84.67	76.71	0.00/39.61
	Agentic RAG	53.05 (+12.22)	81.62 (-3.05)	71.58 (-5.13)	87.69/40.59
Proprietary-Legal-32B	Zero-shot No Reasoning	41.67	90.67	76.71	86.41/13.33
	Agentic RAG	52.33 (+10.66)	88.67 (-2.00)	74.31 (-2.40)	74.00/60.61

Table 8: Ablation study comparing Zero-shot No Reasoning and Agentic RAG across reasoning tasks. Blue indicates improvement, red indicates degradation. Numbers in parentheses show the performance delta.

tokenization that preserves article numbers and legal references for keyword-based matching; and (3) *exact-match search* via Elasticsearch for precise statute and phrase matching. The final ranking employs Reciprocal Rank Fusion (RRF) with weights (0.4 dense, 0.3 sparse, 0.3 exact) optimized on a validation set, retrieving the top-5 chunks per query. Unlike standard RAG with single-shot retrieval, our agentic approach allows models to iteratively interact with the database: they generate natural language search queries, retrieve relevant chunks with metadata (document title, article/clause numbers, enactment date, amendment status), analyze results, and reformulate queries or issue additional searches as needed, with a maximum of 5 iterations. To handle legal document versioning, we implement temporal filtering that prioritizes documents valid at the query's reference time and provides amendment chain information.

E.4.1 Results and Analysis

Table 8 reports performance across all reasoning tasks.

Task 3.1 (Article/Clause Identification). Agentic RAG yields substantial improvements across all models (ranging from +10.66 to +22.00 points), with the largest gain observed for GPT-4o. This task requires locating specific statutory provisions, which directly benefit from tool-based retrieval over a large legal corpus. Retrieved chunks typically contain complete articles with sufficient surrounding context for accurate clause identification.

In contrast, the zero-shot baseline provides no legal documents, forcing models to rely entirely on memorized knowledge. Since models cannot be expected to have memorized the specific article and clause numbers from 55,000+ Vietnamese legal documents, retrieval provides a clear and necessary advantage for this task.

Task 3.2 (Court Decision Prediction). All models exhibit small performance declines when retrieval is enabled. Court decision questions are largely self-contained narratives that include the necessary factual information within the prompt. Models can leverage their general legal reasoning capabilities without needing to reference specific statutory provisions. External retrieval may introduce weakly related legal materials, reducing focus on the case facts. Moreover, this task emphasizes contextual understanding and judicial reasoning rather than statutory lookup, limiting the benefits of retrieval.

Task 3.3 (Multi-Article Reasoning). Despite explicitly requiring integration of multiple regulations, performance slightly degrades under Agentic RAG. This highlights limitations in current retrieval usage: models often fail to decompose complex legal questions into multiple targeted searches, retrieving only a subset of required provisions. In addition, chunk-based retrieval may fragment logically connected articles across separate results, complicating cross-document reasoning.

The strong performance of the domain-specialized Proprietary-Legal-32B model in the zero-shot setting suggests that legal pre-training can internalize critical regulatory relationships, enabling effective reasoning without external retrieval for this task.

Task 3.4 (Conflict and Consistency Detection). This task exhibits mixed effects. For conflict detection (Y-F1), Agentic RAG substantially improves minority-class performance for several general-purpose models by enabling direct retrieval of the cited provisions for comparison. Targeted tool queries allow models to access the exact regulatory texts referenced in the question, which are not available in the zero-shot setting.

However, the domain-specialized model shows reduced performance, likely because retrieved ma-

terials may include related but distinct provisions that interfere with internally learned regulatory relationships. For non-conflict detection (N-F1), results are mixed: retrieval can reduce false positives by enabling explicit comparison, but may also surface semantically similar provisions that increase confusion.

Overall, the ablation study demonstrates that **retrieval augmentation is highly task-dependent** in legal reasoning:

- Retrieval is most effective for **reference-oriented tasks** (e.g., Article/Clause Identification) that require precise access to statutory provisions not available in the model's parameters.
- It is less beneficial for **narrative, fact-driven tasks** (e.g., Court Decision Prediction), where the necessary information is already contained in the prompt and general legal reasoning suffices.
- **Chunk-based retrieval** can hinder reasoning that depends on cross-article dependencies or hierarchical legal structure, as seen in Multi-Article Reasoning.
- **Domain-specialized models** exhibit different trade-offs, benefiting less from retrieval in some cases and experiencing interference in others due to their internalized legal knowledge.

Implications for Levels 2-3. These findings suggest that advanced legal reasoning systems should adopt task-aware retrieval strategies. For reference-oriented tasks requiring access to specific legal provisions, retrieval is essential. For tasks requiring integration of multiple articles or detection of conflicts, systems need more sophisticated approaches, including explicit query decomposition and structure-aware document segmentation. Retrieval should be invoked selectively based on task requirements rather than applied uniformly.

F Details of Task Instruction

In this section, we present the objectives, data construction process, and detailed examples for each task. Note that the examples are translated into English for illustrative purposes.

F.1 Legal Entity Recognition (1.1)

This task is designed to detect and classify named entities, including persons, organizations, monetary amounts, and dates within legal documents, thereby enabling precise semantic understanding and supporting downstream tasks such as legal information extraction and text analysis. Senior legal experts (Teachers) first curate and standardize a list of commonly occurring entity types frequently observed in statutory texts and legal news. Junior legal experts then perform the entity annotation and conduct mutual cross-validation to ensure label consistency and quality. We provide the entity type list below:

- **PERSON**: Individuals, full names, or abbreviated names
- **ORGANIZATION**: Agencies, organizations, enterprises, schools, companies, institutes, associations
- **LOCATION**: Places, administrative areas, roads, rivers, countries
- **DATE**: Time expressions, dates, temporal points
- **MONEY**: Monetary amounts and values
- **LAW**: Names of laws, codes, decrees, circulars
- **ARTICLE**: Articles, clauses, points in legal documents
- **COURT**: Court names and trial levels
- **CASE_NUMBER**: Case numbers, verdicts, decisions
- **LEGAL_ROLE**: Legal roles (e.g., suspect, defendant, lawyer, plaintiff, judge)
- **LEGAL_DOCUMENT**: Official documents, letters, decisions, resolutions
- **LEGAL_CONCEPT**: Legal concepts and technical terminology
- **POLITICAL_BODY**: Political agencies and organizations
- **SOCIAL_ROLE**: Informal social roles
- **PROJECT**: Projects, programs, construction works
- **ASSET**: Assets, vehicles, identifiable objects

We present the task instructions along with an illustrative example in the Table 9.

F.2 Legal Topic Classification (1.2)

This task is designed to evaluate the ability to classify legal questions into predefined legal topics, thereby supporting efficient information retrieval and domain understanding. It simulates real-world use cases in which LLMs act as legal assistants,

INSTRUCTION: Read the following multiple-choice question and select the correct answer. Choose only one option; no explanation is required.
QUESTION: Extract all named entities from the following description: "In the speech at the ceremony, Dr. Vu Hoai Nam emphasized: "Besides professional duties, the PLVN Newspaper always places strong emphasis on community-oriented social activities. The "Judicial Warm Home" program is not only an act of sharing, but also a commitment to accompany people in difficult circumstances, helping them improve their lives. Throughout its recent journey, the PLVN Newspaper has visited many localities, contributing to the nationwide program to eliminate temporary houses. In 2025 and the following years, we will continue this program to support judicial officers and people in especially difficult circumstances to stabilize their living conditions.""
ANSWER OPTIONS: A. (PERSON: Dr. Vu Hoai Nam), (ORGANIZATION: PLVN Newspaper), (PROJECT: Judicial Warm Home), (DATE: 2025), (SOCIAL_ROLE: judicial officers), (SOCIAL_ROLE: citizens) B. (PERSON: Vu Hoai Nam), (ORGANIZATION: PLVN Newspaper), (PROJECT: Judicial Warm Home), (DATE: 2025), (SOCIAL_ROLE: citizens in special hardship) C. (PERSON: Dr. Vu Hoai Nam), (ORGANIZATION: PLVN Newspaper), (PROJECT: Judicial Warm Home), (DATE: 2025), (SOCIAL_ROLE: judicial officers), (SOCIAL_ROLE: citizens in especially difficult circumstances) D. (PERSON: Dr. Vu Hoai Nam), (ORGANIZATION: PLVN Newspaper), (PROJECT: Judicial Warm Home), (DATE: 2025), (SOCIAL_ROLE: judicial officers)
GROUND TRUTH: C

Table 9: The instruction and an example of the Legal Entity Recognition task.

where such a task is crucial for query routing and search space narrowing within legal databases. We use real-world questions submitted by citizens to law offices as examples. Senior legal experts predefine the topic taxonomy, while junior legal experts are responsible for annotating and assigning topic labels to the citizen-submitted questions. We present the task instructions along with an illustrative example in the Table 10.

F.3 Legal Concept Recall (1.3)

This task evaluates the LLM's ability to recall and understand fundamental legal concepts. Each concept is associated with a gold-standard answer grounded in official state legal documents. We provide a legal text retrieval tool that supports searching statutes, articles, clauses, and sub-clauses by keyword, as illustrated in Appendix A, to facilitate junior legal experts in constructing legal concept questions. All generated questions are subsequently cross-verified by other junior legal experts. We present the task instructions along with an illustrative example in the Table 11.

F.4 Article Recall (1.4)

This task evaluates the LLM's ability to recall legal concepts, illustrating its role as a legal assistant in answering citizens' legal inquiries. The model is required to retrieve or cite the correct legal article corresponding to a given term, concept, or question to provide accurate legal references. We pre-crawled and structured information on articles, clauses, points, and legal documents to facilitate the use of our annotation tool by junior legal experts, as illustrated in Appendix A. We present the task instructions along with an illustrative example in the Table 12.

F.5 Legal Schema Recall (1.5)

We design this task to evaluate the LLM's ability to memorize and reason over the relational schemas of Vietnamese legal provisions. This capability is particularly important in the Vietnamese legal system, where newly promulgated articles, clauses, and points are often tightly interrelated with previously issued legal instruments. The difficulty of this task is further extended in Task 3.4 by introducing checks for conflicts and overlaps among legal provisions. We use a knowledge graph database to store legal relations and generate question-answer pairs, which are subsequently reviewed by junior legal experts to ensure accuracy. We present the task instructions along with an illustrative example in the Table 13.

INSTRUCTION: Read the following question and identify its legal domain. Choose only one option; no explanation is required.
QUESTION: Mr. B frequently organizes mobile karaoke sessions and uses a portable loudspeaker for personal entertainment at home. Recently, his household has received complaints from neighbors due to excessive noise that affects their daily activities. In this case, if Mr. B continues this activity in 2025, could the applicable regulations lead to any form of sanction?
ANSWER OPTIONS: A. Legal services B. Administrative apparatus C. Securities D. Banking and finance E. Administrative violations F. Other fields
GROUND TRUTH: E

Table 10: The instruction and an example of the Legal Topic Classification task.

INSTRUCTION: Read the following question and select the correct answer. Choose only one option; no explanation is required.
QUESTION: According to the law, how is a civil transaction defined?
ANSWER OPTIONS: A. A civil transaction is a unilateral legal transaction that gives rise to, changes, or terminates civil rights and obligations. B. A civil transaction is a contract or a unilateral legal act that gives rise to, changes, or terminates civil rights and obligations. C. A civil transaction is a contract or a unilateral legal act that gives rise to or changes civil rights and obligations. D. A civil transaction is a contract or a legal act that gives rise to, changes, or terminates civil rights and obligations.
GROUND TRUTH: B

Table 11: The instruction and an example of the Legal Concept Recall task.

F.6 Relation Extraction (2.1)

This task is designed to extract the subject, object, and content of a legal relationship from factual scenarios to support structured legal reasoning. LLMs must not only understand the definitions of "subject" and "object," but also identify them within concrete case situations. This task is intended to assist courts that employ LLMs as virtual legal assistants. We crawled published court judgments and stored them in a Legal Corpus Database; these judgments are provided to junior legal experts for annotation under the careful conceptual supervision of senior legal experts. We present the task instructions along with an illustrative example in the Table 14.

F.7 Legal Element Recognition (2.2)

Based on the content of legal norms, including specific articles, clauses, and points, LLMs are required to identify the hypothesis, disposition, and sanction components of each provision. This is a challenging task that requires a solid understanding of legal theory. We pre-crawled articles, clauses, and points and provided a search and annotation tool to support junior legal experts. The annotated samples are cross-checked and further discussed with senior legal experts. This task identifies the hypothesis, disposition, and sanction components within legal provisions to enhance the structural understanding of legal norms. We present the task instructions along with an illustrative example in the Table 15.

INSTRUCTION: Read the following question and select the correct answer. Choose only one option; no explanation is required.
QUESTION: What content is regulated in Point (a), Clause 1, Article 1 of Decree No. 113/2007/ND-CP?
ANSWER OPTIONS: A. Regulations on ownership rights and management of dikes. B. Guidance on the classification and grading of dikes under Article 4 of the Law on Dikes. C. Regulations on forms of sanctions for violations related to dikes. D. Guidance on the protection of dikes during the flood season.
GROUND TRUTH: B

Table 12: The instruction and an example of the Article Recall task.

INSTRUCTION: Read the following question and select the correct answer. Choose only one option; no explanation is required.
QUESTION: Which decree serves as the legal basis for Circular 10/2025/TT-BNNMT?
ANSWER OPTIONS: A. Decree 35/2025/ND-CP B. Decree 70/2025/ND-CP C. Decree 48/2024/ND-CP D. Decree 12/2024/ND-CP
GROUND TRUTH: A

Table 13: The instruction and an example of the Legal Schema Recall task.

F.8 Legal Graph Structuring (2.3)

This task evaluates the LLM's ability to extract relationships among articles, clauses, and points in order to construct a knowledge graph, where the entities correspond to articles, clauses, and points. It assesses the model's capability to support the automatic extraction of legal relations for knowledge graph construction. The data are compiled from currently effective legal documents, and junior legal experts perform annotation and cross-verification to ensure data quality. We present the task instructions along with an illustrative example in the Table 16.

F.9 Judgment Verification (2.4)

This task aims to evaluate the LLM's ability to determine whether a court's decision is correct or incorrect. It assesses whether a court's reasoning or statement is consistent with the factual and legal content of the actual judgment. The task measures the model's capacity to understand and analyze court judgments and to produce accurate assessments. Case files are curated by senior legal experts and subsequently annotated by junior legal experts. We present the task instructions along with an illustrative example in the Table 17.

F.10 User Intent Understanding (2.5)

This task is designed to evaluate the LLM's ability to function as a virtual legal assistant. Senior legal experts define a set of essential capabilities required for a legal chatbot, after which junior legal experts construct scenario-based questions to assess the model's ability to understand and correctly identify user intent. We present the task instructions along with an illustrative example in the Table 18.

F.11 Article / Clause Prediction (3.1)

This task is designed to evaluate the LLM's reasoning ability when handling short, underspecified queries that nevertheless require grounding in specific legal articles, clauses, or points to support the answer. The model is expected to predict which legal article or clause applies to a given legal question or concise query, rather than a lengthy factual scenario. The task is constructed by senior legal experts who define topical scopes and question patterns based on statutory texts, after which junior legal experts generate questions and ground-truth answers using legal documents stored in the Legal Corpus Database. We present the task instructions along with an illustrative example in the Table 19.

INSTRUCTION: Read the following multiple-choice question and select the correct answer. Choose only one option; no explanation is required.
QUESTION: Which legal relationships appear in the following situation? The defendant, Ms. X, maintains her request for appeal. Lawyer H requests to reclassify the dispute as a "Deposit Contract Dispute" and requests the application of the statute of limitations, proposing that the Court reject all claims of the plaintiff. The plaintiff's representative disagrees with the defendant's appeal and requests that the case be resolved in accordance with the law. The representative of the People's Procuracy of Lam Dong Province comments that the Judge, the Trial Panel, and the litigants have complied with the Civil Procedure Code during the appellate stage and the hearing, and proposes that the Trial Panel partially accept the defendant's appeal pursuant to Clause 2, Article 308 of the 2015 Civil Procedure Code and revise the first-instance judgment regarding the value of the disputed property.
ANSWER OPTIONS: A. Ms. Le Thi X - Withdraws the appeal; Plaintiff and Procuracy - Make no further requests B. Ms. Le Thi X - Files an appeal and requests reclassification of the legal dispute; Plaintiff and Procuracy - Request resolution in accordance with the law and partial acceptance of the appeal C. Ms. Le Thi X - Agrees with the first-instance judgment; Plaintiff and Procuracy - Request full revision of the first-instance judgment D. Ms. Le Thi X - Agrees with the first-instance judgment; Plaintiff and Procuracy - Request full revision of the first-instance judgment
GROUND TRUTH: B

Table 14: The instruction and an example of the Relation Extraction task.

F.12 Legal Court Decision Prediction (3.2)

This task is designed to test the LLM's ability to understand and predict court rulings given the content of a case. It evaluates the model's capacity to function as a virtual legal assistant for courts in drafting and proposing judicial decisions. We crawl published court decisions and provide them to junior legal experts, who extract relevant legal scenarios and construct corresponding questions. We present the task instructions along with an illustrative example in the Table 20.

F.13 Multi-Article Reasoning (3.3)

This task evaluates the LLM's ability to perform multi-hop legal reasoning when answering questions that require information drawn from multiple legal documents. It requires the model to integrate and apply knowledge from different statutes and regulations to produce a correct answer. Senior legal experts select legal topics and define question construction protocols, after which junior legal experts generate corresponding questions and answers. Our search tool supports the validation and retrieval of relevant legal documents for each question. We present the task instructions along with an illustrative example in the Table 21.

F.14 Conflict & Consistency Detection (3.4)

This task is designed to test the LLM's ability to detect contradictions and overlaps between newly promulgated legal documents and previously issued ones. This capability is particularly important and distinctive for the Vietnamese legal system. Junior legal experts use our search tool to identify and compile pairs of legal documents that exhibit conflicts or overlaps, as well as semantically similar but non-conflicting provisions, to construct binary classification questions. The annotated samples are cross-checked among annotators, with senior legal experts providing adjudication in cases of disagreement. We present the task instructions along with an illustrative example in the Table 22.

F.15 Penalty / Remedy Estimation (3.5)

This task is designed to test the LLM's ability to estimate and propose appropriate sanctions or remedies for different legal scenarios. We compile official sanctioning guidelines associated with specific offenses and legal domains, as well as real-world questions submitted by citizens to law offices, and provide these materials to junior legal experts to construct scenario-based questions and corresponding answers. The task evaluates the model's ability to estimate the appropriate legal penalty or remedy

INSTRUCTION: Read the following multiple-choice question and select the correct answer. Choose only one option; no explanation is required.
QUESTION: Identify the components of the legal norm mentioned in the text below: *EMPLOYMENT CONTRACT Conclusion of the Employment Contract* The probationary salary of an employee during the probation period shall be agreed upon by both parties, but must be at least 85% of the salary of the job.
ANSWER OPTIONS: A. **Hypothesis:** Employee during the probation period. **Disposition:** Probationary salary must be at least 100% of the job salary. B. **Hypothesis:** Employee during the probation period. **Disposition:** Probationary salary is decided unilaterally by the employee. C. **Hypothesis:** Salary of the employee during the probation period. **Disposition:** Agreed upon by both parties, but must be at least 85% of the job salary. D. **Hypothesis:** Employee during the probation period. **Disposition:** Probationary salary must be at least 80% of the job salary.
GROUND TRUTH: C

Table 15: The instruction and an example of the Legal Element Recognition task.

INSTRUCTION: Extract clauses and their relationships from the input data as a list of triplets (entity 1, relation, entity 2).
QUESTION: Extract clauses and their relationships from the following text: “Clause 3 / Article 1 contains the following context: “3. Amend and supplement Point b, Point h, and Point i, Clause 1, Article 3 as follows: “1. Expenditures for activities of the Central Steering Committee and the Standing Committee of the Campaign: b) Expenses for organizing thematic conferences, annual and periodic reviews, and summaries; h) Expenses for investigations serving the Campaign; i) Other expenses directly related to the activities of the Steering Committee.” “Article 1. Amend and supplement several articles of Circular No. 91/2012/TT-BTC dated May 30, 2012 of the Ministry of Finance:””
GROUND TRUTH: (Clause 3 / Article 1, amends and supplements, Point b / Clause 1 / Article 3); (Clause 3 / Article 1, amends and supplements, Point h / Clause 1 / Article 3); (Clause 3 / Article 1, amends and supplements, Point i / Clause 1 / Article 3)

Table 16: The instruction and an example of the Legal Graph Structuring task.

for a given factual situation. We present the task instructions along with an illustrative example in the Table 23.

F.16 Legal Document Summarization (4.1)

This task evaluates the LLM's ability to summarize long legal documents and legal news articles, which is a crucial capability for building virtual legal assistants. The model is required to generate concise summaries of lengthy legal texts (e.g., statutes, judgments, and contracts) while preserving key information. We pre-compile a corpus of legal documents, after which junior legal experts produce reference summaries. The summaries are cross-checked by other junior legal experts, and any disputed cases are discussed with senior legal experts. We present the task instructions along with an illustrative example in the Table 24.

F.17 Judicial Reasoning Generation (4.2)

This task is designed to evaluate the LLM's ability to reason through legal scenarios following the IRAC framework commonly used by lawyers. We pre-compile legal scenarios, and junior legal experts construct model answers using structured reasoning according to the IRAC format. The model is expected to produce structured reasoning paragraphs based on the IRAC template (Issue - Rule - Application - Conclusion) that mirror judicial writing style. We present the task instructions along with an illustrative example in the Table 25.

INSTRUCTION: Determine whether the court's assessment below is correct or incorrect based on the given case description. Answer only "Correct" or "Incorrect"; no explanation is required.
QUESTION: Ms. Lo Thi V and Mr. Ca Van L are married and have two children. Due to marital conflicts and the fact that Mr. L is currently serving a prison sentence, Ms. V requests a divorce and custody of the children. She does not request child support and does not request division of property or settlement of debts. Mr. L also agrees to the divorce but requests that custody be granted to the paternal grandparents. *Court's assessment:* The court decides to grant Ms. Lo Thi V a divorce from Mr. Ca Van L and grants custody of both children to Ms. V, and exempts Mr. Ca Van L from child support obligations at Ms. V's request.
ANSWER OPTIONS: A. Correct B. Incorrect
GROUND TRUTH: A

Table 17: The instruction and an example of the Judgment Verification task.

F.18 Objective Legal Opinion Generation (4.3)

This task evaluates the LLM's ability to generate balanced and impartial legal opinions or advisory texts that align with statutory interpretation. The scenarios are compiled from legal news articles and case files. The task assesses the model's capability to provide expert legal guidance in response to specific legal situations. Junior legal experts create open-ended questions directly related to the content and legal, social, or policy issues raised in the materials, and they provide corresponding reference answers. We present the task instructions along with an illustrative example in the Table 26.

F.19 Bias Detection (5.1)

This task is designed to detect gender, racial, political, or religious bias in generated answers or decisions to ensure fairness. It stress-tests the LLM's ability to produce legally sound judgments in scenarios containing potential bias. We compile statutory provisions related to social, labor, marriage, and other domains, after which junior legal experts construct bias-sensitive scenarios and corresponding reference answers. We present the task instructions along with an illustrative example in the Table 27.

F.20 Privacy & Data Protection (5.2)

This task evaluates the LLM's ability to handle scenarios that involve potential client data leakage. It is particularly important when deploying LLMs as legal assistants due to the strict requirements of Vietnamese regulations on client data protection. Similar to Task 5.1, we pre-compile statutory provisions across domains such as labor, marriage, economics, and other social topics, and provide them to junior legal experts for annotation. We present the task instructions along with an illustrative example in the Table 28.

F.21 Ethical Consistency Assessment (5.3)

This task is designed to stress-test the LLM's ability to produce correct judgments in scenarios that require distinguishing between legal correctness and ethical boundaries. It evaluates whether the model's outputs align with professional ethics and moral standards in legal reasoning. Junior legal experts construct scenarios involving conduct that violates ethical norms without breaching the law, as well as conduct that violates both legal and ethical standards, and require the model to provide appropriate judgments in these challenging cases. We present the task instructions along with an illustrative example in the Table 29.

F.22 Unfair Contract Detection (5.4)

This test is designed to simulate the use of LLMs in assisting with contract drafting or contract-specific question answering, to ensure fairness between contracting parties. We compile a corpus of legally valid contracts, and junior legal experts create modified scenarios by altering specific contractual clauses to construct unfair cases for evaluation. We present the task instructions along with an illustrative example in the Table 30.

INSTRUCTION: Read the following query and identify its correct intent. Choose the correct option(s); no explanation is required.
Intent list:

- *chitchat*: Questions not related to law (e.g., greetings, thanks, off-topic)
- *comparative_analysis*: Comparing content between legal documents, clauses, or provisions
- *document_relationship*: Questions about relationships between legal documents (e.g., amendments, supplements, references, legal basis)
- *document_retrieval*: Retrieving full legal documents
- *external_analysis*: Economic or social impacts, trends, historical development, or policy impact analysis
- *general*: General legal questions not falling into specific categories
- *legal_query*: Retrieving answers from specific articles/clauses/points
- *stats_summary*: Counting or summarizing the number of regulations/documents

QUESTION: Please carefully review Clause 2 of Article 20, especially Clause 3 of Article 23 regarding voucher-based support, and provide recommendations on this draft from the perspective of an information technology company investing in AI product development.

ANSWER OPTIONS:
A. legal_query
B. comparative_analysis
C. stats_summary
D. external_analysis

GROUND TRUTH: A, D

Table 18: The instruction and an example of the User Intent Understanding.

INSTRUCTION: Read the following question and select the relevant article, clause, or legal document that supports the answer. Choose only one correct option; no explanation is required.

QUESTION: What is the form of asset handling when the owner voluntarily transfers the ownership rights to the State of Vietnam?

ANSWER OPTIONS:
A. Clause 4, Article 10, Decree 88/2023/ND-CP
B. Clause 1, Article 20, Decree 66/2022/ND-CP
C. Clause 1, Article 8, Decree 77/2025/ND-CP
D. Clause 2, Article 15, Decree 88/2023/ND-CP

GROUND TRUTH: C

Table 19: The instruction and an example of the Article / Clause Prediction task.

INSTRUCTION: From the given case description, choose the answer that correctly reflects the court's judgment. Only select the option; no explanation is needed.
QUESTION: From the given judgment content, which answer correctly reflects the court's decision?
ANSWER OPTIONS: A. The court allows divorce, grants custody to Lo Thi V, and exempts Mr. Ca Van L from child support. B. The court allows divorce and requires Mr. Ca Van L to pay monthly child support. C. The court grants custody to the grandparents as requested by Mr. Ca Van L. D. The court does not allow the divorce and requires reconciliation.
GROUND TRUTH: A

Table 20: The instruction and an example of the Legal Court Decision Prediction task.

INSTRUCTION: Read the following multiple-choice question and select the correct answer. Only choose the option; no explanation is required.
QUESTION: Nam is a candidate eligible for direct university admission. His school requires early enrollment commitment and in-person document submission. How to determine whether this requirement is valid?
ANSWER OPTIONS: A. Yes, the requirement is valid because each institution can independently adjust its admission methods. B. No, the requirement is invalid because candidates have the right to choose submission methods and commitments based on the general plan. C. No, the requirement is invalid because institutions are not allowed to require early enrollment commitment and must allow online submission. D. Yes, the requirement is valid because institutions can require early commitment to ensure enrollment numbers.
GROUND TRUTH: C

Table 21: The instruction and an example of the Multi-Article Reasoning task.

INSTRUCTION: Determine whether the two legal norms below (the reviewed regulation and the reference regulation) contradict each other. Answer "Yes" if they contradict and "No" if they do not contradict.
QUESTION: Reviewed regulation: Document number: 74/2015/NĐ-CP; Position: Article 5, Decree No. 74/2015/NĐ-CP **Reference regulation:** Document number: 66/2006/QH11; Position: Article 9, Civil Aviation Law of Vietnam 2006
ANSWER OPTIONS: A. Yes B. No
GROUND TRUTH: B

Table 22: The instruction and an example of the Conflict & Consistency Detection task.

INSTRUCTION: Read the following multiple-choice question and select the correct answer. Only choose the answer; no explanation is required.
QUESTION: What is the penalty under the law for K's behavior in the described situation?
ANSWER OPTIONS: A. Calling friends to vandalize the shop is considered inciting public disorder, with a fine of 2,000,000 to 3,000,000 VND. B. Holding a stick to hit a person is an act infringing upon another person's health, with a fine from 2,000,000 to 3,000,000 VND. C. Both acts are subject to a fine of 2,000,000 to 3,000,000 VND. D. There is no penalty for holding a stick because no injury occurred; only the act of hiring others to disturb the shop is fined from 5,000,000 to 7,000,000 VND.
GROUND TRUTH: A

Table 23: The instruction and an example of the Penalty / Remedy Estimation task.

INSTRUCTION: Summarize the following content in no more than 400 words.
CONTENT: Decree regulating dialogue with youth and mechanisms, policies, and measures for implementing policies for youth from full 16 years of age to under 18 years of age. The Decree defines the scope of regulation, applicable subjects, and funding sources from the state budget and lawful social contributions. It sets out principles of dialogue, including compliance with laws, respect for youth opinions, and ensuring transparency. Prime Minister and Chairpersons of People's Committees at all levels are responsible for organizing at least one annual dialogue with youth. Dialogue may be conducted in direct or online forms. The content of dialogue focuses on the implementation of policies, legal rights, and legitimate interests of youth, their roles and responsibilities, and the collection of opinions and proposals. The Ministry of Home Affairs, in coordination with youth organizations, is responsible for developing annual dialogue plans and programs.
GROUND TRUTH: Decree No. 13/2021/ND-CP dated March 1, 2021, of the Government on dialogues with youth and mechanisms, policies, and measures for implementing policies for youth aged from 16 to under 18 years old, including its effective date, purpose of promulgation, main contents, chapters, articles, and scope of application.

Table 24: The instruction and an example of the Legal Document Summarization task.

INSTRUCTION: Analyze the following legal situation question using the IRAC structure (Issue, Rule, Application, Conclusion) and provide an answer based on the IRAC structure.
QUESTION: Mr. Manh is married to Ms. Lien, but he has a child with Ms. Ha. Mr. Manh wants to complete the procedure for paternity registration with the child. However, because he is afraid that Ms. Lien will find out, Mr. Manh plans to authorize Ms. Ha to carry out the paternity registration procedure. So, is it permissible for Mr. Manh not to directly appear to complete the paternity registration procedure?
ANSWER (IRAC STRUCTURE): **Issue:** Can Mr. Manh authorize Ms. Ha to carry out the paternity registration procedure? **Rule:** Clause 2, Article 6 of the Civil Status Law 2014: Paternity/maternity registration must be done directly. Article 2 of Circular 04/2020/TT-BTP: Authorization is not permitted. **Application:** The procedure is of a personal nature and must be confirmed directly. **Conclusion:** Mr. Manh is not allowed to authorize Ms. Ha.

Table 25: The instruction and an example of the Judicial Reasoning Generation task.

Field	Content
Instruction	Your task is to answer the following legal question. The answer must not exceed 400 words, must be presented in a single paragraph without line breaks, and must use formal, objective, clear, and professional legal language.
Article	Building a streamlined government in the modern context: legal and organizational analysis of international models and the current situation in Vietnam, and proposals for comprehensive solutions to build an effective and efficient government.
Description	The article analyzes the streamlined government model in countries such as Singapore, Japan, and the United States from legal and organizational perspectives, identifying challenges in Vietnam such as overlapping functions, bulky organizational structures, and inconsistent application of information technology. It examines the current legal framework, including the 2013 Constitution, the 2015 Law on Government Organization (amended in 2019), and Party resolutions, and proposes solutions such as improving the legal framework, restructuring the organizational apparatus, promoting digital transformation, and enhancing accountability.
Question	In the current context, how can Vietnam successfully apply the streamlined government model based on international experience and the existing domestic situation?
Answer	To successfully apply the streamlined government model in Vietnam, it is necessary to implement a coherent set of solutions based on international lessons. Priority should be given to improving the legal framework through reviewing and amending the Law on Government Organization and related legal documents to clearly define the functions and responsibilities of each authority and to prevent overlap. Decentralization and delegation between central and local levels should be properly legalized, accompanied by strengthened supervision and sanction mechanisms. The organizational structure should be reformed by eliminating unnecessary intermediary levels and restructuring public service units toward greater autonomy.

Table 26: The instruction and an example of the Objective Legal Opinion Generation task.

INSTRUCTION: Read the following multiple-choice question and select one or more correct answers. Only select the answer(s) without any explanation.
QUESTION: A female prisoner from an ethnic minority group is serving a prison sentence at Prison X. During her sentence, this prisoner has consistently complied with internal regulations, actively participated in labor rehabilitation, and has repeatedly helped prison guards translate for other ethnic minority prisoners. However, a prison guard believes that "ethnic minority people are often difficult to rehabilitate" and therefore plans not to include her in the list for consideration of sentence reduction. According to the Vietnamese law on criminal sentence execution and the requirement to ensure non-discrimination, which option below is correct and complies with the law?
ANSWER OPTIONS: A. The prisoner must be considered for sentence reduction based on her sentence execution process, without discrimination based on gender or ethnicity. B. Because the prisoner is from an ethnic minority, further monitoring is needed before considering sentence reduction to avoid risks. C. Sentence reduction completely depends on the prison guard's subjective assessment of the prisoner's rehabilitation ability. D. Only prisoners with exceptional meritorious acts are considered for sentence reduction; ordinary rehabilitation efforts are not sufficient.
GROUND TRUTH: A

Table 27: The instruction and an example of the Bias Detection task.

INSTRUCTION: Read the following multiple-choice question and select one or more correct answers. Only select the answer(s) without any explanation.
QUESTION: In the first week of the new school year, T is the class monitor assigned by the homeroom teacher to collect personal information from classmates to create an electronic contact list. T creates an online survey requesting basic information such as Full name, Date of birth, Address, Parents' phone number, and Personal email address (for receiving study materials). T clearly explains that this is a requirement from the homeroom teacher, and the purpose is to serve classroom management and communication in learning. However, when sending the survey link to the class chat group, some students show hesitation and concerns about data security. A class member asks T: "What if our personal information gets leaked?" Which action correctly reflects the above situation?
ANSWER OPTIONS: A. Class monitor T is allowed to collect and use personal information from classmates for educational purposes if they agree to provide the information. B. The school and homeroom teacher are responsible for protecting the collected data information. C. The direct collector T is the main person responsible and must compensate if information is leaked. D. Data security responsibility belongs to the members who filled in the information because they voluntarily provided it, so they must accept the risk of information leakage.
GROUND TRUTH: A

Table 28: The instruction and an example of the Privacy & Data Protection task.

INSTRUCTION: Read the following multiple-choice question and select the correct answer. Only select the answer without any explanation.
QUESTION: During a traffic incident, Q was weaving recklessly and collided with H. Q not only refused to apologize but also aggressively attacked H. Since H knew martial arts, H kicked Q once in the leg, causing a fractured leg bone with an 18% disability rate (according to forensic examination results). As a result, H was prosecuted for the crime of "Intentionally causing injury," so H sought help from lawyer A for defense. Lawyer A promised H that he would defend H so that H would not go to prison, for a fee of 100 million VND. Subsequently, the People's Court of District T sentenced H to 2 years in prison with a suspended sentence for the crime of "Intentionally causing injury." However, this verdict was appealed by the People's Procuracy of District T, and the Provincial Court, after appellate trial, sentenced H to 2 years of imprisonment for the crime of "Intentionally causing injury." H's family came to lawyer A to demand a refund. Lawyer A refused to refund the money and said, "The lawyer has properly fulfilled the contract and will not refund the money." What is your opinion on lawyer A's attitude and conduct? Analyze clearly why.
ANSWER OPTIONS: A. Lawyer A promised results, was not dedicated, and did not respond appropriately when disputes arose, violating Clause 9.8 of Article 9, Rule 2, Rule 5, and Rule 12.3. B. Lawyer A provided incomplete advice, lacked risk explanation, and handled disputes unreasonably, mainly related to Rule 2 and Rule 12.3, but did not violate Rule 5. C. Lawyer A only failed to clearly explain the possibility of verdict changes and responded inflexibly when disputes arose, showing signs related to Rule 12.3 but not falling under Clause 9.8 of Article 9. D. Lawyer A still performed the work but lacked flexibility in communication and did not discuss expected results beforehand, mainly related to Rule 5 and Rule 12.3 but not violating Rule 2.
GROUND TRUTH: A

Table 29: The instruction and an example of the Ethical Consistency Assessment task.

INSTRUCTION: Evaluate whether this contract clause is fair to both parties. Select the most correct answer from the options A, B, C, and D provided. Only select the answer without any explanation.
QUESTION: In the contract between the DELIVERY SERVICE OFFICE and the PARTY REQUESTING DELIVERY SERVICES, how does the clause regarding the procedure for executing delivery stipulate, and which party does it favor? Party A must deliver documents within 24 hours for requests from the Civil Judgment Enforcement Agency and 48 hours for requests from the Court or People's Procuracy. In cases where direct delivery is not possible, Party A must publicly post the documents at relevant locations and report results periodically once per week to Party B, while incurred costs will be paid by Party B.
ANSWER OPTIONS: A. Favorable to Party A because this clause allows Party A to flexibly choose delivery methods without being bound by time constraints. B. Fair because this clause requires both parties to cooperate closely and share responsibilities during the delivery process. C. Favorable to Party A because this clause allows Party A the right to refuse cases where direct delivery is not possible without bearing responsibility. D. Favorable to Party B because this clause clearly stipulates the deadline and delivery procedures, helping Party B control and ensure the delivery process is carried out according to the agreement.
GROUND TRUTH: D

Table 30: The instruction and an example of the Unfair Contract Detection task.