

ViLegalLM: Language Models for Vietnamese Legal Text

Truong-Phuc Nguyen, Quy-Nhan Nguyen, Minh-Tien Nguyen*

Faculty of Information Technology,

Hung Yen University of Technology and Education, Hung Yen, Vietnam

{nguyentruongphuc_12421tn, nguyenquynhan, tienm}@utehy.edu.vn

Abstract

We present **ViLegalLM**, comprising **ViLegalBERT** and **ViLegalQwen**, the first suite of Vietnamese pretrained language models for legal text understanding and generation. It includes one encoder-only model (ViLegalBERT, 135M parameters) and two decoder-only models (ViLegalQwen2.5-1.5B-Base and ViLegalQwen3-1.7B-Base), all continually pre-trained on a newly curated 16GB Vietnamese legal corpus, significantly larger than previous work. To mitigate data scarcity, we construct three synthetic datasets using LLM-based generation and hard negative mining for True/False QA, Multiple Choice QA, and Natural Language Inference. We establish state-of-the-art results among open-source models on four main Vietnamese legal downstream tasks spanning ten benchmarks, demonstrating that continual pretraining from base models consistently outperforms instruction-tuned adaptation. Source codes, corpus, datasets, and model checkpoints are publicly available at <https://github.com/ntphuc149/ViLegalLM>.

1 Introduction

Self-supervised pretraining has revolutionized NLP, enabling strong performance across diverse tasks. Encoder-only models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), and decoder-only LLMs like GPT (Radford et al.), LLaMA (Grattafiori et al., 2024), and Qwen (Yang et al., 2025) leverage large-scale unlabeled corpora with self-supervised objectives, such as Masked Language Modeling (MLM) (Devlin et al., 2019) and Causal Language Modeling (CLM) (Radford et al.), to learn robust representations for downstream tasks. This paradigm extends to Vietnamese with models like PhoBERT (Nguyen and Tuan Nguyen, 2020), ViDeBERTa (Tran et al., 2023), and PhoGPT (Nguyen et al., 2023a).

However, general-purpose models often underperform in specialized domains due to distinct linguistic characteristics and domain-specific vocabulary (Lee et al., 2020; Chalkidis et al., 2020). In response to this, domain-adapted models have been developed for different domains, such as for biomedicine in both English and Vietnamese (e.g., BioBERT (Lee et al., 2020), BioELECTRA (Kanakarajan et al., 2021), ViHealthBERT (Minh et al., 2022), ViPubMedDeBERTa (Tran-Tien et al., 2023)), demonstrating superior performance on NLP downstream tasks such as named entity recognition, natural language inference, and question answering.

A particularly challenging and high-stakes domain is legal text processing, characterized by long sentences, specialized terminology, and intricate syntax, requiring dedicated models for tasks such as judgment prediction, question answering, and contract analysis (Zhong et al., 2020; Hendrycks et al., 2021). While Legal-BERT (Chalkidis et al., 2020) and SaulLM-7B (Colombo et al., 2024) have shown strong results for English, Vietnamese legal text processing remains underserved. Table 1 summarizes Vietnamese pre-trained models by domain: while multiple domains have dedicated models via continual pretraining on large-scale corpora, legal text has limited options: VNLawBERT (Chau et al., 2020) (179M, 320MB text) and qwen3-legal-pretrain (Le et al., 2025) (1.7B/4B, 2.52GB from instruction-tuned checkpoints). These existing approaches face key limitations: relatively small corpora compared to general-domain scales (320MB-2.52GB vs. 20-482GB for general domain), potential catastrophic forgetting of instruction-following capabilities under continuous pre-training (Jindal et al., 2024), and lack of comprehensive solutions providing both encoder and decoder architectures with large-scale curated legal corpora.

We introduce **ViLegalLM**, comprising **ViLegalBERT** (an encoder-only model) and **ViLegalQwen**

*Corresponding author.

Table 1: Pre-trained models for Vietnamese. Rep, gen denote representation, generation, respectively.

Pretrained models	#Params	Max length	Domain	Corpus size	Task
PhoBERT _{base} (Nguyen and Tuan Nguyen, 2020)	135M	256	General	20GB	rep
ViDeBERTa _{base} (Tran et al., 2023)	86M	512	General	138GB	rep
VNLawBERT (Chau et al., 2020)	179M	512	Legal	0.32GB	rep
ViSoBERT (Nguyen et al., 2023b)	79M	512	Social	1GB	rep
ViHealthBERT _{word-base} (Minh et al., 2022)	135M	256	Healthcare	≈0.84GB+25M sent (OSCAR)	rep
ViPubmedDeBERTa _{base} (Tran-Tien et al., 2023)	79M	512	Biomedical	≈18.12GB	rep
PhoGPT-4B (Nguyen et al., 2023a)	3.70B	8192	General	482GB	gen
qwen3-1.7b-legal-pretrain (Le et al., 2025)	1.70B	4096	Legal	≈2.52GB	gen
qwen3-4b-legal-pretrain (Le et al., 2025)	4.00B	4096	Legal	≈2.52GB	gen
ViLegalBERT (Ours)	135M	256	Legal	16GB	rep
ViLegalQwen2.5-1.5B-Base (Ours)	1.54B	2048	Legal	16GB	gen
ViLegalQwen3-1.7B-Base (Ours)	1.72B	4096	Legal	16GB	gen

(two decoder-only models), continually pretrained on a newly curated 16GB Vietnamese legal corpus. We address data scarcity through three synthetic datasets and conduct comprehensive evaluation across four main Vietnamese legal tasks spanning ten benchmarks. Our main contributions are:

- We introduce ViLegalLM comprising three domain-adapted models (encoder-only and decoder-only) continually pretrained from base models on a newly curated 16GB Vietnamese legal corpus, demonstrating better performance over recent baselines.
- We build three synthetic datasets and provide empirical analysis across ten benchmarks, offering practical baselines for the community to build upon for more complex Vietnamese legal reasoning tasks.
- We publicly release corpora, model checkpoints, synthetic datasets, and source code to advance Vietnamese legal NLP research.

2 Related work

Legal text processing in high-resource languages

Legal text processing in high-resource languages has developed comprehensive resources for both pre-training and downstream task evaluation. For English, MultiLegalPile (Niklaus et al., 2024) provides 689GB covering 24 languages, LegalBench (Guha et al., 2023) offers extensive evaluation tasks. For Chinese, large-scale corpora have enabled benchmarks such as LawBench (Fei et al., 2024) and CMDL (Huang et al., 2024). Continual pre-training on such domain-specific corpora has emerged as the most effective approach for model adaptation (Gururangan et al., 2020; Ling et al., 2025). For English, encoder-only models like LEGAL-BERT (Chalkidis et al., 2020) and

LexLM (Chalkidis et al., 2023) employed MLM, while decoder-only SaulLM-7B (Colombo et al., 2024) used CLM on large-scale legal texts. For Chinese, Lawformer (Xiao et al., 2021) adapted encoders for long documents, and LawGPT (Zhou et al., 2024) applied CLM on legal corpora. These successes in domain adaptation motivate the development of legal language models for Vietnamese.

LegalAI in Vietnamese Vietnamese legal text processing, however, remains in early development. VNLawBERT (Chau et al., 2020), the first pre-trained model for Vietnamese legal text, continually pre-trained mBERT on a 320MB corpus and achieved strong performance on answer selection tasks. The annual ALQAC competition (Do et al., 2024) has provided high-quality datasets for information retrieval and question answering. Most recently, the VLSP LegalSLM challenge (Le et al., 2025) introduced two models (1.7B and 4B) that continually pre-trained instruction-tuned Qwen3 models on 2.52GB of legal texts (144,737 documents and news articles). These models achieved promising results on multiple-choice legal knowledge, natural language inference, and syllogism reasoning. However, existing approaches face key limitations: limited corpus sizes (320MB-2.52GB) compared to resource-rich languages, and domain adaptation from instruction-tuned rather than base models. Addressing these limitations, we introduce ViLegalLM, a comprehensive suite comprising one encoder and two decoder models continually pre-trained from base checkpoints on 16GB of high-quality Vietnamese legal text.

3 ViLegalLM

We present three Vietnamese legal language models: ViLegalBERT and ViLegalQwen (2.5-1.5B-Base, 3-1.7B-Base), continually pretrained on our

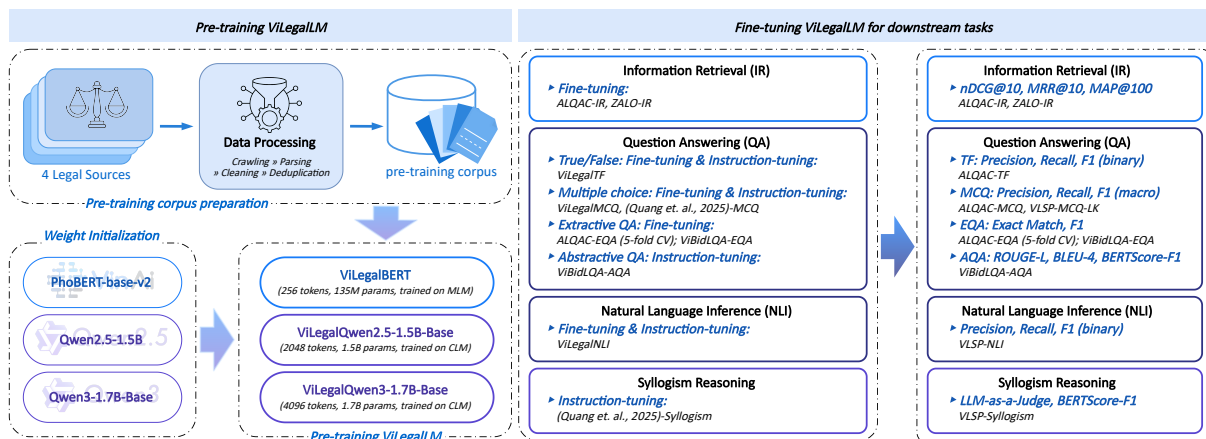


Figure 1: Overview of pre-training and fine-tuning ViLegalLM.

16GB legal corpus. Figure 1 illustrates the pre-training and fine-tuning pipeline.

3.1 Pre-training Corpus

We constructed a 16GB Vietnamese legal corpus from four publicly available legal repositories¹ with multi-stage preprocessing to ensure corpus quality for continual pretraining. We successfully crawled 1.39M documents from 1.40M identified URLs across the four repositories. Retrieved HTML documents underwent content extraction using BeautifulSoup with source-specific parsing rules to extract legal text, yielding 1.14M successfully parsed documents in total 19GB of Vietnamese legal text.

The 19GB raw corpus underwent a multi-stage preprocessing pipeline. First, we applied NFKC normalization to convert all text to canonical Unicode representation. Second, we performed comprehensive text cleaning to remove HTML tags, escape sequences, formatting codes, URLs, emails, non-linguistic symbols, and normalize punctuation and whitespace. Third, we applied FastText language identification (Joulin et al., 2017) to retain only Vietnamese documents (confidence ≥ 0.95), reducing the corpus from 19GB to 18GB. Finally, we employed MinHashLSH deduplication (Broder, 1997; Indyk and Motwani, 1998) with threshold 0.8 to remove near-duplicates while preserving legitimately similar legal texts, reducing the corpus to 16GB. For ViLegalBERT training, we additionally word-segmented the corpus using Vietnamese toolkit PyVi to match PhoBERT’s tokenization.

¹TVPL: <https://thuvienphapluat.vn>; VBPL: <https://vbpl.vn>; LuatVietnam: <https://luatvietnam.vn>; LawNet: <https://lawnet.vn>;

3.2 Model Architecture

We developed three models through continual pre-training from strong foundations for Vietnamese NLP. ViLegalBERT (135M parameters), built on PhoBERT-base-v2 (Nguyen and Tuan Nguyen, 2020), retained the maximum sequence length of 256 tokens to ensure stable continual pre-training while providing sufficient context for legal document segments. ViLegalQwen2.5-1.5B-Base (1.54B parameters), built on Qwen2.5-1.5B (Team, 2024) with 2,048-token context, and ViLegalQwen3-1.7B-Base (1.72B parameters), built on Qwen3-1.7B-Base (Yang et al., 2025) with extended 4,096-token context. Training both models allows comparing the effect of model capacity and context length across legal reasoning tasks.

3.3 Continual Pre-training

We employed continual pretraining to adapt foundation models to the Vietnamese legal domain. In particular, ViLegalBERT used MLM (Devlin et al., 2019) for training by masking 15% of tokens to maximize log-likelihood given bidirectional context. Following RoBERTa (Liu et al., 2019), we excluded Next Sentence Prediction as it provided minimal pretraining benefits. ViLegalQwen models employed CLM (Radford et al.), predicting next tokens autoregressively for generative capabilities. All models were trained with the AdamW (Loshchilov and Hutter, 2017) optimizer, gradient clipping, and early stopping based on validation loss monitoring. Complete training configurations are provided in Appendix A.

4 Downstream Vietnamese Legal Datasets

Table 2: Statistics of Vietnamese legal benchmark datasets across four main tasks. #train, #val., #test: training, validation, and test sets; IR, QA, NLI, Syllogism: Information Retrieval, Question Answering, Natural Language Inference, and Syllogism Reasoning; TF, MCQ(-LK), EQA, AQA: True/False, Multiple Choice (Legal Knowledge), Extractive QA (span-based extraction), and Abstractive QA (answer generation); [●]: existing synthetic datasets; [◆]: our synthetic datasets; †: 5-fold cross-validation.

Tasks	Sub tasks	#train		#val.		#test	
		Name	Sample	Name	Sample	Name	Sample
IR	-	ALQAC-IR	647	ALQAC-IR	82	ALQAC-IR	82
		ZALO-IR	2996	ZALO-IR	100	ZALO-IR	100
QA	TF	ViLegalTF [◆]	13032	ViLegalTF [◆]	388	ALQAC-TF	387
	MCQ	ViLegalMCQ [◆]	14920	ViLegalMCQ [◆]	300	ALQAC-MCQ	286
	MCQ-LK	Quang et al.-MCQ [●]	63717	Quang et al.-MCQ [●]	146	VLSP-MCQ-LK	146
	EQA	ALQAC-EQA†	416	-	-	ALQAC-EQA†	104
	EQA	ViBidLQA-EQA	4798	ViBidLQA-EQA	500	ViBidLQA-EQA	1000
	AQA	ViBidLQA-AQA	4798	ViBidLQA-AQA	500	ViBidLQA-AQA	1000
NLI	-	ViLegalNLI [◆]	7660	ViLegalNLI [◆]	150	VLSP-NLI	150
Syllogism	-	Quang et al.-Syllogism [●]	60293	Quang et al.-Syllogism [●]	144	VLSP-Syllogism	144

This section describes the Vietnamese legal datasets. We leveraged 10 existing benchmarks and developed 3 synthetic datasets to address data scarcity. Table 2 shows complete dataset statistics.

4.1 Existing Datasets

ALQAC ALQAC (Do et al., 2024) provides comprehensive Vietnamese legal benchmarks. For IR, we split the dataset into 647/82/82 train/validation/test samples. For QA, ALQAC only provides labeled training data (TF: 387, MCQ: 286, EQA: 520). We, therefore, used ALQAC’s TF and MCQ training sets as test sets and synthetic ViLegalTF and ViLegalMCQ for training. For EQA, we employed 5-fold cross-validation (416 train/104 test per fold) due to the difficulty of generating synthetic extractive questions. The ALQAC corpus grounds all three synthetic datasets, ensuring legal accuracy and domain relevance.

ZALO ZALO AI Challenge 2021² provides a Vietnamese legal IR benchmark (3,271 documents, 61,425 articles, 3,196 queries). ZALO data was used for IR evaluation and, together with ALQAC, to generate ViLegalNLI via hard negative mining.

ViBidLQA Vietnamese Bidding Law Question Answering (Nguyen et al., 2025a; Phung et al., 2025) provides 6,298 expert-verified QA pairs for EQA and AQA, serving as the large-scale benchmark focused on generative legal reasoning. Following prior work, we used train/test splits and created a validation set from training data.

VLSP VLSP 2025 Legal SLM competition (Le et al., 2025) provides three tasks: MCQ-LK, NLI,

and Syllogism Reasoning, each with ≈ 150 public test samples. We used synthetic training data from Bosch@AI_Team (Quang et al., 2025) for MCQ-LK (63,863 samples) and Syllogism (60,437 samples), and our ViLegalNLI (7,810 samples) for NLI. Validation sets matched test sizes.

4.2 Our Synthetic Datasets

To address the shortage of labeled training data for True/False, Multiple Choice, and Natural Language Inference tasks in the Vietnamese legal domain, we developed three synthetic datasets using task-appropriate methodologies, all grounded in the verified ALQAC legal corpus.

ViLegalTF and ViLegalMCQ We generated TF and MCQ datasets through a 4-stage pipeline, as illustrated in Figure 2.

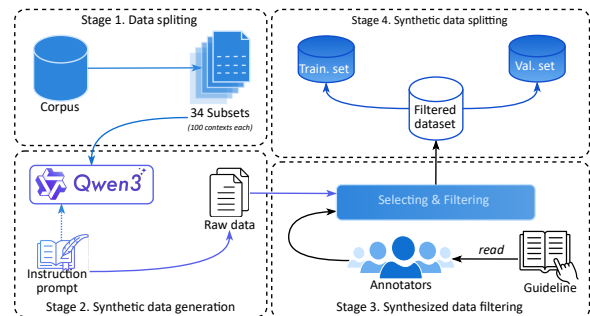


Figure 2: ViLegalTF&ViLegalMCQ data construction.

First, the ALQAC legal corpus was divided into 34 subsets of 100 articles each for parallel generation. Second, we employed Qwen3-8B with 1-shot prompting and thinking mode to generate questions with difficulty stratification, legal terminology grounding, and structured formatting, mo-

²<https://challenge.zalo.ai/>

tivated by its demonstrated instruction-following capability on Vietnamese legal QA tasks (Nguyen et al., 2025b; Trieu et al., 2025). Generated questions included model-provided explanations for quality verification. Third, five trained annotators who are university graduates with legal domain familiarity performed systematic quality control through a verification protocol assessing legal context comprehension, question clarity, answer validity, label correctness, and generation completeness. Finally, filtered samples were split into stratified train/validation sets matching test distributions.

For ViLegalTF, we generated 4 questions per legal article at hierarchical difficulty levels (Recognition, Comprehension, Application, High Application). From 13,600 raw samples, human filtering yielded 13,420 qualified samples (1.32% rejection rate, primarily from incomplete generation), with final split 13,032 train/388 validation. For ViLegalMCQ, we generated 5 questions per legal citation across three difficulty bands (easy, medium, difficult). From 17,000 raw samples, human filtering produced 15,220 qualified samples, with final split 14,920 train/300 validation. To assess annotation reliability, we measured full-consensus agreement (the proportion of samples where all five annotators unanimously agreed on the accept/reject decision) on a random subset of 100 samples per dataset, yielding 0.71 for ViLegalTF and 0.68 for ViLegalMCQ, both indicating strong consensus.

ViLegalNLI We created an NLI dataset through hard negative mining from ALQAC and ZALO benchmarks, as shown in Figure 3.

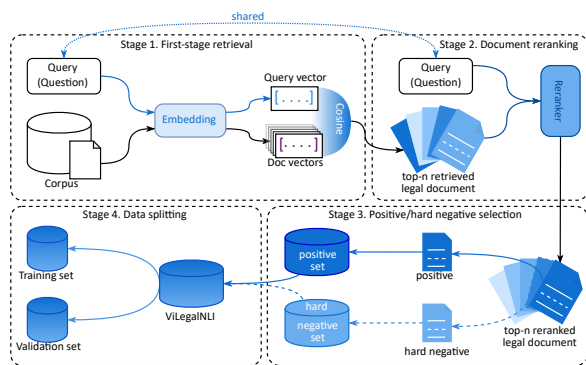


Figure 3: ViLegalNLI dataset creation pipeline.

We leveraged state-of-the-art Vietnamese embedding and reranking models (Nguyen Nho Trung, 2025)³, achieving top performance on the ZALO

³https://hf.co/AITeamVN/Vietnamese_Embedding;
https://hf.co/AITeamVN/Vietnamese_Reranker

Legal Text Retrieval benchmark. For each legal question, we (1) retrieved top-10 documents, (2) reranked via cross-encoder for accurate query-document interaction scoring, and (3) selected ground-truth documents as positives and high-ranking non-relevant documents as hard negatives. When the ground truth ranked first, the second-ranked document became the hard negative; otherwise, the top-ranked document served as the hard negative. This created challenging examples where models must distinguish semantically similar but factually incorrect legal citations. Final split: 7,660 train/150 validation, tested on VLSP-NLI.

5 Experiments

We evaluated ViLegalLM against existing Vietnamese language models across ten benchmarks using task-specific fine-tuning and standard metrics for each task type.

5.1 Evaluation Metrics

We employed standard metrics for each task.

Information Retrieval We evaluated ranking quality using nDCG@k (Järvelin and Kekäläinen, 2002), MRR@k, and MAP@k. We set k=10 for nDCG and MRR to focus on top-ranked results, while MAP used k=100 for stable measurement across larger result sets.

Question Answering For True/False and Multiple Choice QA, we reported Precision, Recall, and F1-Score with binary and macro-averaging, respectively. For Extractive QA, we followed SQuAD (Rajpurkar et al., 2016) and reported Exact Match (EM) and token-level F1-Score. For Abstractive QA, we employed standard generation metrics: ROUGE (Lin, 2004) (ROUGE-L), BLEU (Papineni et al., 2002) (BLEU-4), and BERTScore-F1 (Zhang et al., 2019).

Natural Language Inference For binary legal NLI (whether legal context has sufficient information to answer a question), we employed Precision, Recall, and F1-Score with binary averaging.

Syllogism Reasoning For syllogistic legal reasoning requiring structured argumentation, we adopt LLM-as-a-Judge (Zheng et al., 2023) as the primary metric, following the evaluation protocol of the VLSP-Syllogism benchmark (Le et al., 2025). Specifically, GPT-4.1⁴ evaluates each gen-

⁴GPT-4.1: gpt-4.1-2025-04-14

erated syllogism on four binary criteria: (C1) *Relevance* (the answer addresses the question), (C2) *Legal Citation* (accurate and specific statutory references), (C3) *Reasoning Accuracy* (legally valid reasoning grounded in cited provisions), and (C4) *Conclusion Accuracy* (the conclusion follows from rules and facts). Each criterion contributes 0.25 points, yielding a per-sample score $s \in [0, 1]$, and we report the mean score \bar{s} as a final result. We also report BERTScore-F1 (Zhang et al., 2019) as a complementary semantic similarity metric.

5.2 Baselines

We compared our domain-adapted models against existing Vietnamese language models across encoder and decoder architectures.

Encoder-only models PhoBERT-base and PhoBERT-base-v2 (Nguyen and Tuan Nguyen, 2020) (135M parameters) represented a foundation for Vietnamese representation learning, pretrained on 20GB and 140GB of Vietnamese text, respectively. VNLawBERT (Chau et al., 2020) (179M parameters), continually pretrained from mBERT on 320MB Vietnamese legal corpus, served as legal domain baseline for encoder-only models.

Decoder-only models We evaluated against base models from the Qwen family: Qwen2-1.5B (Yang et al., 2024), Qwen2.5-1.5B (Team, 2024) (1.54B), and Qwen3-1.7B-Base (Yang et al., 2025) (1.72B), represented different generations of decoder architectures. qwen3-1.7b-legal-pretrain (Le et al., 2025) (1.72B) represented the latest legal LLM for Vietnamese, providing a parameter-matched comparison with our ViLegalQwen3-1.7B-Base. We omitted qwen3-4b-legal-pretrain and PhoGPT-4B due to computational constraints. We additionally evaluated *gpt-4o-mini* in a 0-shot setting (same instruction format as Appendix B.2, no in-context examples) as a closed-source upper-bound reference across applicable tasks.

5.3 Experimental Setup

We employed two fine-tuning paradigms based on model architecture: discriminative fine-tuning for encoder-only models and instruction tuning with QLoRA (Detmers et al., 2023) for decoder-only models. All models were fine-tuned on NVIDIA P100 16GB GPUs. Complete experimental configurations are provided in Appendix B.

6 Results and Discussion

We evaluated ViLegalLM against existing Vietnamese language models across four main tasks spanning ten benchmarks. In all result tables, **bold** and underlined values indicate the best and second-best performance within each open-source parameter group, respectively; *italic* values denote closed-source model scores (reported as upper-bound references where applicable).

6.1 Information Retrieval

We presented information retrieval results on ALQAC-IR and ZALO-IR in Table 3.

Table 3: IR results on ALQAC-IR and ZALO.

Models	nDCG@10	MRR@10	MAP@100
<i>ALQAC</i>			
PhoBERT-base	0.5094	0.4613	0.4720
PhoBERT-base-v2	0.6279	0.5710	0.5779
VNLawBERT	<u>0.6581</u>	<u>0.6053</u>	<u>0.6081</u>
ViLegalBERT (ours)	0.6786	0.6248	0.6304
<i>ZALO</i>			
PhoBERT-base	0.5533	0.5165	0.5188
PhoBERT-base-v2	0.5936	0.5541	0.5597
VNLawBERT	0.6020	<u>0.5550</u>	<u>0.5609</u>
ViLegalBERT (ours)	0.6300	0.5878	0.5912

ViLegalBERT achieves SOTA results on both benchmarks across all metrics, outperforming VN-LawBERT despite having fewer parameters. Compared to the base checkpoint PhoBERT-base-v2, continual legal pretraining yields substantial improvements across both benchmarks, demonstrating the effectiveness of domain-specific pretraining even when building from strong general-domain foundations. These consistent gains validate the effectiveness of large-scale legal corpus pretraining for dense retrieval in the Vietnamese legal domain.

6.2 Question Answering

We presented QA results in Table 4 across four subtasks: True/False (TF), Multiple Choice (MCQ), Extractive (EQA), and Abstractive (AQA).

Across TF and MCQ tasks, our ViLegalLM models demonstrated consistent improvements over their respective base checkpoints. For TF QA, ViLegalBERT achieved the best F1 score among encoder-only models, leading in recall while maintaining competitive precision. Among decoder-only models, ViLegalQwen2.5-1.5B-Base achieved the best performance in the 1.5B parameter group across most metrics, while ViLegalQwen3-1.7B-Base reached the best results across all three met-

Table 4: Question Answering results across four subtasks. VLQ: ViLegalQwen. Pre/Rec/F1: Precision/Recall/F1-Score, EM: Exact Match, RG-L/BL-4/BS-F1: ROUGE-L/BLEU-4/BERTScore-F1, mac: macro-averaging. –: not evaluated on this subtask.

Models	True/False			Multiple Choice						Extractive QA				Abstractive QA		
	ALQAC-TF			ALQAC-MCQ			VLSP-MCQ-LK			ALQAC-EQA		ViBidLQA-EQA		ViBidLQA-AQA		
	Pre	Rec	F1	Pre _{mac}	Rec _{mac}	F1 _{mac}	Pre _{mac}	Rec _{mac}	F1 _{mac}	EM	F1	EM	F1	RG-L	BL-4	BS-F1
PhoBERT-base	67.26	77.16	71.87	62.03	62.77	62.05	55.75	53.18	53.34	37.25	64.46	49.34	77.45	–	–	–
PhoBERT-base-v2	76.88	67.51	71.89	62.96	63.70	62.50	57.65	54.46	55.70	40.58	68.01	50.66	78.59	–	–	–
VNLawBERT	72.91	75.13	74.00	61.00	61.12	60.80	56.87	52.35	54.19	38.82	65.85	48.49	76.32	–	–	–
ViLegalBERT (ours)	70.97	78.17	74.40	62.76	63.71	62.83	58.05	56.30	56.53	41.17	65.92	50.19	78.63	–	–	–
Qwen2-1.5B	85.05	86.84	85.94	82.19	81.05	81.42	68.02	53.62	58.15	–	–	–	–	72.44	49.00	89.68
Qwen2.5-1.5B	74.47	92.11	82.35	84.80	84.05	84.37	65.05	52.34	56.54	–	–	–	–	73.02	49.12	89.74
VLQ2.5-1.5B-Base (ours)	87.31	90.53	88.89	85.66	84.53	84.96	65.24	54.54	58.39	–	–	–	–	73.45	49.90	89.91
Qwen3-1.7B-Base	90.27	87.89	89.07	85.64	84.68	85.03	67.84	62.80	64.98	–	–	–	–	73.81	51.18	90.24
qwen3-1.7b-legal-pretrain	89.62	86.32	87.94	87.96	88.19	88.00	66.95	60.88	63.32	–	–	–	–	74.84	51.49	90.32
VLQ3-1.7B-Base (ours)	90.62	91.58	91.10	89.22	88.81	88.92	70.12	64.00	66.54	–	–	–	–	74.49	52.11	90.43
<i>gpt-4o-mini</i>	89.86	97.89	93.70	90.83	91.58	91.15	69.05	51.16	58.17	–	–	–	–	67.06	40.46	86.85

rics, surpassing both base Qwen3 and qwen3-1.7b-legal-pretrain. Continual pretraining from base models with larger legal corpora proves more effective than adapting instruction-tuned models.

For MCQ, ViLegalLM models demonstrated consistent improvements across both context-based and knowledge-intensive variants. On ALQAC-MCQ with legal context, our models achieved the best or second-best performance within their parameter groups. The knowledge-intensive VLSP-MCQ-LK (without context) proved substantially more challenging, yet ViLegalBERT and ViLegalQwen variants achieved the best F1 scores in their respective groups. Despite lower absolute scores compared to context-based MCQ, these results highlight the difficulty of pure legal knowledge recall versus context-based reasoning.

For extractive QA, ViLegalBERT achieved the best EM on ALQAC-EQA with limited training data, suggesting legal domain adaptation benefits precise boundary detection in low-resource settings. On ViBidLQA-EQA with more training data, PhoBERT-base-v2 surpassed ViLegalBERT’s EM while ViLegalBERT maintained an F1 advantage, indicating that abundant supervision partially compensates for lack of domain adaptation, though legal pretraining still benefits semantic matching.

For abstractive QA, both ViLegalQwen variants demonstrated strong performance. ViLegalQwen2.5-1.5B-Base achieved the best scores in the 1.5B parameter group across all three metrics, while ViLegalQwen3-1.7B-Base achieved the best scores in semantic similarity and n-gram precision within the 1.7B parameter group. While qwen3-1.7b-legal-pretrain achieved marginally higher ROUGE-L, our model excelled in BERTScore and BLEU-4.

Regarding the closed-source upper bound, *gpt-4o-mini* achieves strong performance on TF and context-based MCQ, performing substantially better than open-source models; however, these gains likely reflect the instruction-following capability of a much larger closed-source model on context-provided tasks rather than domain-specific legal knowledge. On knowledge-intensive MCQ-LK, where context is absent, *gpt-4o-mini* scores lower, falling below our ViLegalQwen3-1.7B-Base, demonstrating that small domain-adapted models are more competitive on tasks requiring Vietnamese legal knowledge.

6.3 Natural Language Inference

We presented NLI results on VLSP-NLI in Table 5.

Table 5: NLI results on VLSP-NLI. (Pre: Precision, Rec: Recall, F1: F1-Score)

Models	Pre	Rec	F1
PhoBERT-base	59.26	21.33	31.37
PhoBERT-base-v2	55.26	28.00	37.17
VNLawBERT	62.50	13.33	21.98
ViLegalBERT (ours)	72.28	97.33	82.95
Qwen2-1.5B	92.86	86.67	89.66
Qwen2.5-1.5B	100.00	80.00	88.89
ViLegalQwen2.5-1.5B-Base (ours)	84.90	100.00	91.84
Qwen3-1.7B-Base	94.00	97.33	95.64
qwen3-1.7b-legal-pretrain	97.44	97.22	97.24
ViLegalQwen3-1.7B-Base (ours)	95.75	100.00	97.83
<i>gpt-4o-mini</i>	100.00	86.67	92.86

Consistent with QA results, ViLegalLM models demonstrated substantial improvements on NLI. ViLegalQwen3-1.7B-Base achieved the best F1 score in its parameter group, performing better than both qwen3-1.7b-legal-pretrain and base Qwen3, with both ViLegalQwen variants achieving perfect recall. This validates base-model continual pretraining with larger legal corpora over instruction-

tuned adaptation across diverse reasoning tasks. Consistent with MCQ-LK results, *gpt-4o-mini* scores lower than all 1.7B open-source models on NLI, further confirming that domain-adapted models at this scale hold a clear advantage on Vietnamese legal knowledge tasks.

Among encoder-only models, ViLegalBERT achieved the best F1 score, outperforming VN-LawBERT despite having fewer parameters. This improvement suggests that large-scale legal pre-training enables distinguishing hard negative citations, which are semantically similar but factually distinct legal provisions that challenge models lacking domain-specific knowledge. VNLawBERT’s lower performance likely reflects the combination of multilingual initialization and limited domain data, whereas ViLegalBERT benefited from monolingual initialization and a larger legal corpus.

6.4 Syllogism Reasoning

We presented legal syllogism reasoning results on the VLSP-Syllogism benchmark in Table 6.

Table 6: Syllogism results on VLSP-Syllogism. (BS-F1: BERTScore-F1, LLM-Judge: LLM-as-a-Judge)

Models	BS-F1	LLM-Judge
Qwen2-1.5B	76.19	0.2639
Qwen2.5-1.5B	76.89	0.2656
ViLegalQwen2.5-1.5B-Base (ours)	76.63	0.2674
Qwen3-1.7B-Base	76.69	0.2760
qwen3-1.7b-legal-pretrain	76.80	0.2899
ViLegalQwen3-1.7B-Base (ours)	77.50	0.3038
<i>gpt-4o-mini</i>	78.63	0.5069

ViLegalQwen3-1.7B-Base achieved the best performance across all metrics, performing better than both qwen3-1.7b-legal-pretrain and base Qwen3. Per-criterion breakdown (Appendix C) reveals that ViLegalQwen3 surpasses qwen3-1.7b-legal-pretrain particularly on Legal Citation and Conclusion Accuracy, suggesting that base-model continual pretraining may offer modest improvements on Legal Citation and Conclusion Accuracy over instruction-tuned adaptation. In contrast, ViLegalQwen2.5-1.5B-Base showed mixed results, achieving the best LLM-Judge score while showing slight decreases in BERTScore-F1 compared to Qwen2.5-1.5B, illustrating that surface-level similarity metrics may not fully capture improvements in structured legal argumentation. All models score substantially below the *gpt-4o-mini* upper bound, with Legal Citation as the weakest criterion across all small models (see Table 11). This highlights ac-

curate statutory citation as the primary bottleneck for Vietnamese legal syllogism reasoning.

6.5 Error Analysis

6.5.1 NLI

We inspected *gpt-4o-mini* misclassifications on VLSP-NLI, categorized by cross-encoder similarity score (Nguyen Nho Trung, 2025). Errors occur even at high similarity (≥ 0.5), prompting qualitative inspection that identifies one dominant error type. **Over-cautious inference under succinct legal language:** the legal context contains sufficient information to answer the question, yet *gpt-4o-mini* predicts “No” because the terse, clause-enumerative style of Vietnamese statutory text does not match the broader framing of the question. In a representative case (similarity 0.9849), the question asks for the legal consequences of exceeding the deductible interest expense limit; Article 16 of Decree 132/2020/ND-CP directly enumerates the 30% cap and the five-year carry-forward of excess interest, yet *gpt-4o-mini* rejects the context, apparently treating itemised regulatory sub-clauses as insufficient for a holistic “*legal consequence*” query. Domain-adapted ViLegalLM models, trained directly on hard-negative pairs from the same legal corpus, correctly classify this representative case.

6.5.2 Syllogism Reasoning

We analyzed 144 ViLegalQwen3-1.7B-Base outputs using LLM-judge and identified three error types. **E1 (Missing or Wrong Citation):** the model either omits all statutory references entirely, substituting them with generic phrases (e.g., “*pur-suant to the law*”), or cites non-existent decisions, outdated circulars, or misattributed articles, affecting all 139 C2-failing samples. **E2 (Flawed Reasoning):** the model fails to apply legal provisions to the specific scenario facts, omitting procedural deadlines or eligibility conditions (C3 fails in 120 of 144 cases). **E3 (Incorrect Conclusion):** the conclusion does not logically follow from the premises or contains factual errors such as wrong penalty amounts (C4 fails in 125 of 144 cases).

The dominant failure pattern (65.27% of samples) is C1=1, C2=0, C3=0, C4=0, and C3 passes in 60.00% of cases when C2 passes versus only 15.11% when C2 fails, suggesting a citation-to-reasoning cascade. Table 7 illustrates one representative case per error type. In the E1 case, the absence of any statutory citation causes the minor premise to ignore the key legal consequence

Table 7: Representative error cases for each error type. **Bold** text highlights the key difference between reference and generated answers. “...” denotes truncated content. Ref.: reference answer; Gen.: model-generated answer.

Error	Major Premise	Minor Premise	Conclusion
E1	Ref. Pursuant to Section 2, Part 7 of Decree 151/2025/ND-CP and Clause 3, Article 172 of the 2024 Land Law , land users ... must submit extension applications within the prescribed deadline; the dossier comprises ... Extension procedures and processes ...	Mr. A’s land use term has expired; he has a complete extension dossier but submitted it after the prescribed deadline ... Mr. A does not request a new Certificate of Land Use Rights ... Meanwhile, another organization has submitted a written request to change ...	Because Mr. A did not submit the dossier within the prescribed deadline, the right to extension is not guaranteed ... Mr. A is not entitled to an extension and must carry out procedures for change of land use purpose or land recovery.
	Gen. The land use term is determined by the term recorded on the Certificate of Land Use Rights. (<i>no citation</i>)	Mr. A submitted an extension application but was not approved by the district-level People’s Committee. (<i>ignores filing deadline</i>)	Mr. A is not granted an extension but only has the change of term confirmed on the issued certificate. ... (<i>misses legal consequence</i>)
E2	Ref. Pursuant to Decision 642/QD-BNNMT of 2025 and Decree 101/2024/ND-CP , ... if the dossier is incomplete or inconsistent, it shall not be accepted and shall be returned ; ... no more than 20 working days, and may be extended by 10 days for mountainous communes. ...	A domestic economic organization in a mountainous commune submitted an incomplete dossier ... the land use origin has not been confirmed due to missing documents ; the processing time is 20 working days plus an additional 10 working days due to the receiving authority has the right to refuse acceptance and return the dossier due to missing documents; ... in the case of an incomplete dossier, the processing period shall not be counted; therefore the organization must supplement the dossier and resubmit it ...
	Gen. Decision 642/QD-BNNMT and Decree 101/2024/ND-CP provide regulations on cases of incomplete dossiers and processing deadlines. (<i>correct citation, vague content</i>)	The organization submitted a complete dossier but was missing some documents, as confirmed by the People’s Committee at the commune level. (<i>contradictory premise</i>)	The organization is allowed to continue processing the initial land registration dossier within the extended deadline, ensuring its rights and obligations ... (<i>wrong: incomplete dossier cannot be processed</i>)
E3	Ref. Pursuant to Clause 8, Article 6 of Decree 168/2024/ND-CP ... a fine of VND 16,000,000 to VND 18,000,000 shall be imposed. Pursuant to Point c, Clause 16, Article 6 of this Decree the above conduct shall also result in a deduction of 06 points from the driver’s license.	Mr. A drove a motor vehicle causing a traffic accident but did not stop, did not preserve the scene, did not assist the victim, and did not report to the police authority.	Mr. A shall be subject to an administrative fine of VND 16,000,000 to VND 18,000,000 and a deduction of 06 points from his driver’s license pursuant to Decree 168/2024/ND-CP.
	Gen. Decree 168/2024/ND-CP provides clear regulations on penalties for failure to stop, failure to preserve the scene, failing to assist (<i>correct decree, missing article/clause</i>)	Mr. A committed these acts and had 12 points deducted from his driver’s license. (<i>wrong point deduction introduced in minor premise</i>)	Mr. A is subject to administrative penalties pursuant to Article 3 of Decree 168/2024/ND-CP and 12 points deducted ... (<i>wrong article, wrong deduction</i>)

of the missed filing deadline, propagating an incomplete conclusion. In the E2 case, citations are presented but their vague content introduces a self-contradictory minor premise, stating the dossier is both complete and incomplete, which reverses the correct legal outcome. In the E3 case, a missing article and clause number allows an incorrect numerical value (12 instead of 6 deduction points) to enter the minor premise and carry through directly into the final conclusion.

7 Conclusion

We present ViLegalLM, comprising ViLegalBERT, ViLegalQwen2.5-1.5B-Base, and ViLegalQwen3-1.7B-Base, continually pretrained on 16GB legal

text, a newly created Vietnamese legal corpus. We develop three synthetic datasets addressing data scarcity and establish state-of-the-art results among open-source models across ten benchmarks spanning four tasks. Our findings demonstrate that base-model continual pretraining with large-scale legal corpora consistently outperforms instruction-tuned adaptation, domain-adapted models hold a competitive advantage over closed-source models on Vietnamese legal knowledge tasks, and accurate statutory citation remains the primary bottleneck for legal syllogism reasoning. We publicly release pre-training corpus, model checkpoints, synthetic datasets, and source code to advance Vietnamese legal NLP research.

Acknowledgement

We thank the anonymous reviewers for their valuable comments and constructive suggestions that helped improve this work. We thank Mr. Manh Tran-Tien for his contributions to pre-training corpus construction. We also thank the five annotators for their quality verification of the synthetic datasets. Computing resources were partly sponsored by Intelligent Integration Co., Ltd. (INT2), Vietnam. We also thank the organizers of the ALQAC competition, ZALO Legal Text Retrieval challenge, and VLSP LegalSLM competition for providing high-quality verified datasets that served as benchmarks and enabled our synthetic dataset construction. Without these contributions, this research would not have been feasible.

Limitations

Although we have achieved promising results, our work has several limitations. First, computational constraints limit the evaluation to models with up to 1.7B parameters, leaving questions about scaling behavior. Second, encoder-only models use 256-token contexts, potentially restricting performance on longer legal documents. Third, we lack statistical significance testing and human evaluation for generation quality. Fourth, synthetic datasets, while human-filtered and grounded in verified legal content, may not fully capture real-world question complexity. Fifth, error analysis covers only NLI and Syllogism Reasoning; broader qualitative analysis remains for future work.

Ethics Statement

We address several ethical considerations relevant to developing and deploying Vietnamese legal pre-trained language models.

Intended Use and Limitations ViLegalLM is designed to support legal NLP research in Vietnamese, including document retrieval, question answering, and text analysis. These models are not intended to replace professional legal counsel; users should not rely solely on model outputs for consequential legal decisions without qualified professional consultation.

Data and Privacy Our 16GB pretraining corpus was collected exclusively from four publicly available Vietnamese legal repositories (TVPL, VBPL, LuatVietnam, LawNet), containing only official legal texts and publicly accessible materials. No

personally identifiable information was included. Models may inherit biases present in these sources, including temporal bias, regional variations, and domain coverage imbalances.

Risk of Misuse Potential misuse risks include providing legal advice without proper licensing, applying model outputs to consequential decisions without human oversight, or treating outputs as definitive legal interpretations. We notice that the use in high-stakes legal scenarios should be confirmed by expert validation.

Synthetic Data Our synthetic datasets were generated using LLMs and verified by five trained annotators with legal domain familiarity. While grounded in verified ALQAC legal corpus, synthetic data may not fully capture real-world legal complexity. All datasets are publicly released for community scrutiny.

Environmental and Accessibility Training required 144, 168, and 24 GPU hours (A100/H100) for ViLegalBERT, ViLegalQwen2.5, and ViLegalQwen3 respectively. We publicly release all corpus, checkpoints, datasets, and code to promote reproducible research. ViLegalLM is trained on Vietnamese legal texts and may not generalize to other jurisdictions in other languages.

We encourage users and researchers to report any ethical concerns or unintended behaviors observed in ViLegalLM through our public repository.

References

1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.
- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. [LEGAL-BERT: The muppets straight out of law school](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.
- Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. [LeXFiles and LegalLAMA: Facilitating English multinational legal language model development](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.

- Chieu-Nguyen Chau, Truong-Son Nguyen, and Le-Minh Nguyen. 2020. Vnlawbert: A vietnamese legal answer selection approach using bert language model. In *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 298–301. IEEE.
- Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and 1 others. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Dinh-Truong Do, Son T Luu, Trang Pham, Trung Vo, Nguyen-Hoang Chu, Quang-Huy Chu, Cuong Nguyen, Minh Nguyen, An Trieu, Dat Nguyen, and 1 others. 2024. A summary of the alqac 2024 competition. In *2024 16th International Conference on Knowledge and System Engineering (KSE)*, pages 422–427. IEEE.
- Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 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, Miami, Florida, USA. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. **Don’t stop pretraining: Adapt language models to domains and tasks**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. **Cuad: An expert-annotated nlp dataset for legal contract review**. *Preprint, arXiv:2103.06268*.
- Wanhong Huang, Yi Feng, Chuanyi Li, Honghan Wu, Jidong Ge, and Vincent Ng. 2024. **CMDL: A large-scale Chinese multi-defendant legal judgment prediction dataset**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5895–5906, Bangkok, Thailand. Association for Computational Linguistics.
- Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613.
- Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.
- Ishan Jindal, Chandana Badrinath, Pranjal Bharti, Lakkidi Vinay, and Sachin Dev Sharma. 2024. **Balancing continuous pre-training and instruction fine-tuning: Optimizing instruction-following in llms**. *Preprint, arXiv:2410.10739*.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. **Bag of tricks for efficient text classification**. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. **BioELECTRA: pretrained biomedical text encoder using discriminators**. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.
- Anh-Cuong Le, Trong-Chi Duong, Viet-Ha Nguyen, and Thang VQ Le. 2025. **Overview of the Legal-SLM shared task: Evaluating legal reasoning of Vietnamese small language models**. In *Proceedings of the 11th International Workshop on Vietnamese Language and Speech Processing*, pages 147–152, Hanoi, Vietnam. Association for Computational Linguistics.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

- Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhagash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, and 5 others. 2025. [Domain specialization as the key to make large language models disruptive: A comprehensive survey](#). *ACM Comput. Surv.*, 58(3).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Nguyen Minh, Vu Hoang Tran, Vu Hoang, Huy Duc Ta, Trung Huu Bui, and Steven Quoc Hung Truong. 2022. [ViHealthBERT: Pre-trained language models for Vietnamese in health text mining](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 328–337, Marseille, France. European Language Resources Association.
- Dat Quoc Nguyen, Linh The Nguyen, Chi Tran, Dung Ngoc Nguyen, Dinh Phung, and Hung Bui. 2023a. [PhoGPT: Generative Pre-training for Vietnamese](#). *arXiv preprint*, arXiv:2311.02945.
- 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, Online. Association for Computational Linguistics.
- Nam Nguyen, Thang Phan, Duc-Vu Nguyen, and Kiet Nguyen. 2023b. [ViSoBERT: A pre-trained language model for Vietnamese social media text processing](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5191–5207, Singapore. Association for Computational Linguistics.
- Thu-Ha Nguyen, Truong-Phuc Nguyen, and Khang Tran.T. 2025a. [Vibidlqa: Vietnamese bidding legal question answering dataset](#).
- Truong-Phuc Nguyen, Quy-Nhan Nguyen, Manh-Cuong Phan, Chi-Hai Cao, Trinh-Hoai-An Duong, and Minh-Tien Nguyen. 2025b. [Utehy-nlu@alqac 2025: Dynamic weighted ensemble and adaptive reasoning for vietnamese legal text processing](#). In *2025 17th International Conference on Knowledge and System Engineering (KSE)*, pages 1–5.
- Nguyen Van Huy Nguyen Nho Trung, Nguyen Nhat Quang. 2025. [Vietnamese_embedding: Embedding model in vietnamese language](#).
- Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel Ho. 2024. [MultiLegalPile: A 689GB multilingual legal corpus](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15077–15094, Bangkok, Thailand. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Nhu Hai Phung, Chi Thanh Nguyen, Minh-Tien Nguyen, Thu Ha Nguyen, Huu Loi Le, and Truong-Phuc Nguyen. 2025. [A fine-tuning framework based on question, context, and answer relationships for enhancing legal information retrieval](#). *Engineering Applications of Artificial Intelligence*, 159:111570.
- Tran Minh Quang, Nguyen Xuan Phi, Nguyen Van Tai, Phan Minh Toan, and Dang Van Thin. 2025. [Bosch@AI_Team at LegalSML 2025: Vietnamese legal small language with domain adaptation and aspect-based data synthesis](#). In *Proceedings of the 11th International Workshop on Vietnamese Language and Speech Processing*, pages 153–164, Hanoi, Vietnam. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. [Improving language understanding by generative pre-training](#).
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Cong Dao Tran, Nhut Huy Pham, Anh Tuan Nguyen, Truong Son Hy, and Tu Vu. 2023. [ViDeBERTa: A powerful pre-trained language model for Vietnamese](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1071–1078, Dubrovnik, Croatia. Association for Computational Linguistics.
- Manh Tran-Tien, Huu-Loi Le, Dang Nhat Minh, T. Tran Khang, Huy-The Vu, and Nguyen Minh-Tien. 2023. [ViPubMedDeBERTa: A pre-trained model for Vietnamese biomedical text](#). In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 831–840, Hong Kong, China. Association for Computational Linguistics.
- Huy Trieu, Dang-Phuong-Nam Doan, Anh-Kiet Nguyen, Thanh-Thai Nguyen, Thanh-Nghia Vo, Tung Le, and Huy Tien Nguyen. 2025. [Flame reavers@alqac 2025: Integrating learned rankers and llm reasoning in a dynamic hybrid architecture for legal retrieval](#). In *2025 17th International Conference on Knowledge and System Engineering (KSE)*, pages 1–6.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

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

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. [Lawgpt: A chinese legal knowledge-enhanced large language model](#). *Preprint*, arXiv:2406.04614.

A Continual Pretraining Configuration Details

Table 8 provides complete hyperparameter configurations for continual pretraining of all three ViLegalLM models.

ViLegalBERT training uses AdamW optimizer (Loshchilov and Hutter, 2017) with learning rate 2e-5, batch size 64, weight decay 0.01, and linear warmup over 10,000 steps followed by linear decay. Training proceeds for 500,000 planned steps on a single A100 GPU (40GB), converging at 410,000 steps (82% of planned) after 6 days with FP32 precision. ViLegalQwen2.5-1.5B-Base uses learning rate 5e-5, batch size 12 with 4-step gradient accumulation (effective batch size 48), weight

Table 8: Continual pretraining configurations across all models.

Configuration	ViLegalBERT	ViLegalQwen2.5-1.5B	ViLegalQwen3-1.7B
Base Model			
Foundation	PhoBERT-base-v2	Qwen2.5-1.5B	Qwen3-1.7B-Base
Training Data			
Corpus	16GB legal text	16GB legal text	16GB legal text
Max sequence length	256	2,048	4,096
Training Objective			
Objective	MLM (15% mask)	CLM	CLM
Hardware			
GPU	A100 40GB	A100 40GB	H100 80GB
Precision	FP32	BF16	BF16
Batch Configuration			
Batch size	64	12	6
Gradient accumulation	1	4	4
Effective batch size	64	48	24
Tokens per update	~16K	~98K	~98K
Optimization			
Optimizer	AdamW	AdamW (fused)	AdamW (fused)
Learning rate	2e-5	5e-5	5e-5
LR schedule	Linear warmup+decay	Cosine+warmup	Cosine+warmup
Warmup steps	10,000	2,000	2,000
Weight decay	0.01	0.1	0.1
Gradient clipping	1.0	1.0	1.0
Training Progress			
Planned steps	500,000	50,000	50,000
Actual steps	410,000 (82%)	22,000 (44%)	25,500 (51%)
GPU hours	149	175	28
Early Stopping			
Monitoring metric	Validation loss	Validation loss	Validation loss
Patience	5 evaluations	5 evaluations	5 evaluations

decay 0.1, and cosine learning rate schedule with 2,000 warmup steps. Training proceeds for 50,000 planned steps on a single A100 GPU (40GB), converging at 22,000 steps (44%) after 7 days with BF16 precision. ViLegalQwen3-1.7B uses identical hyperparameters with batch size 6 and 4-step gradient accumulation (effective batch size 24), training for 50,000 planned steps on a single H100 GPU (80GB), converging at 25,500 steps (51%) after 1 day with BF16 precision. All models employ gradient clipping at norm 1.0 for training stability and monitor validation loss on held-out legal documents, applying early stopping when validation loss plateaus for five consecutive evaluations.

B Experimental Setup Details

This section describes the fine-tuning methods used for ViLegalLM and baselines. We provide instruction templates for decoder-only models across six tasks requiring generative reasoning: True/False QA, Multiple Choice QA, Multiple Choice Legal Knowledge, Abstractive QA, Natural Language Inference, and Syllogism Reasoning.

B.1 Fine-tuning Paradigms

We employed two fine-tuning paradigms: discriminative fine-tuning for encoder-only models and instruction tuning for decoder-only models to align them with downstream tasks.

Encoder-only models We applied discriminative fine-tuning with task-specific classification heads

for ViLegalBERT and baselines. IR employed a dual-encoder architecture with contrastive learning and hard negative mining (3 negatives per query). QA and Classification tasks used standard classification heads atop the [CLS] token, while EQA employed span-based prediction with start and end position classifiers.

Decoder-only models For ViLegalQwen variants and baselines, we employed instruction tuning with QLoRA (Dettmers et al., 2023), which combines 4-bit quantization with LoRA adapters (rank 16, alpha 32). All tasks followed the Qwen chat template format (Yang et al., 2024), reformulating tasks as instruction-following generation.

B.2 Instruction template for fine-tuning ViLegalQwen

We employed instruction tuning for ViLegalQwen and baseline decoder models across three tasks: Question Answering, Natural Language Inference, and Syllogism Reasoning.

B.2.1 Question Answering

Question Answering encompasses four subtasks: True/False, Multiple Choice, Multiple Choice Legal Knowledge, and Abstractive QA.

True/False We formulated True/False QA as binary classification, where models select between two options (True or False) based on legal context.

```
<|im_start|>system
You are a Vietnamese legal expert with extensive experience in evaluating the accuracy of legal claims based on provided legal documents.<|im_end|>
<|im_start|>user
Legal Text:
Constitution:
When the President is unable to perform duties for an extended period, the Vice President shall act as President. In the event of a vacancy in the office of the President, the Vice President shall act as President until the National Assembly elects a new President.

Question:
In the case where the President is unable to perform duties for a short period of time, the Vice President shall act as President. Is this statement true or false?

Options:
A. True
B. False

Based on the legal text above, determine whether the statement in the question is true or false.<|im_end|>
<|im_start|>assistant
B<|im_end|>
```

Multiple choice Multiple Choice QA requires selecting the best answer from 4 options (A/B/C/D).

```
<|im_start|>system
You are a Vietnamese legal expert with extensive experience in analyzing legal texts and selecting accurate answers to legal questions.<|im_end|>
<|im_start|>user
Legal Text:
Law on Access to Information
Information Citizens Are Entitled to Access
Citizens are entitled to access information held by state agencies, except for information that is restricted from access as prescribed in Article 6 of this Law; and may access information subject to conditions as prescribed in Article 7 of this Law.

Question:
Which article stipulates the cases in which information is restricted from citizen access?

Options:
A. Article 5
B. Article 6
C. Article 7
D. Article 8

Based on the legal text above, please select the correct answer.<|im_end|>
<|im_start|>assistant
B<|im_end|>
```

Multiple choice Legal Knowledge The key difference between MCQ and MCQ-LK is that MCQ-LK requires models to select the correct answer without legal context, relying solely on legal knowledge. We employed step-by-step reasoning with <think> tags to enhance model performance on this challenging variant.

```
<|im_start|>system
You are a Vietnamese legal expert with extensive experience in analyzing legal documents and selecting accurate answers to legal questions.<|im_end|>
<|im_start|>user
Question:
Which of the following activities is regulated by the Law on Geology and Minerals No. 54/2024/QH15?

Options:
A. Fundamental geological investigation, protection of geological resources, extraction and processing of minerals throughout the entire territory of Vietnam
B. Only regulates petroleum and natural mineral water extraction activities
C. Only applies to onshore activities, and does not regulate activities on islands and at sea
D. Only provides regulations on mineral processing not associated with extraction projects

Please provide the correct answer.<|im_end|>
<|im_start|>assistant
<think>
The correct answer is option A because, according to Article 1, the Law comprehensively regulates the investigation, protection, extraction, processing, financial management, and auctioning of mineral extraction rights throughout the territory of Vietnam, including inland areas, islands, internal waters, and maritime zones. The remaining options either emphasize only a partial scope or incorrectly describe the regulated scope of the Law.
</think>
A<|im_end|>
```

Abstractive QA Similar to other text generation tasks, we formatted inputs with task instructions, legal context, and questions. Models learn to generate answers based on the provided legal context.

```
<|im_start|>system
You are a Vietnamese legal expert with many years of experience, specializing in analyzing legal documents and answering questions accurately and in detail.<|im_end|>
<|im_start|>user
Based on the following legal text:
Law on Bidding No. 22/2023/QH15
Article 96. Transitional Provisions
2. Business investment projects that have been approved and for which bidding documents were issued prior to the effective date of this Law shall continue to proceed with investor selection, contract signing, and contract performance management in accordance with the Law on Bidding No. 43/2013/QH13 and its implementing and guiding documents. The Government shall provide detailed regulations on the application of transitional provisions to business investment projects.

Please provide an answer to the following question:

Question:
Which authority is responsible for providing detailed regulations on the application of transitional provisions to business investment projects?<|im_end|>
<|im_start|>assistant
Pursuant to the provisions of Article 96. Clause 2 of the Law on Bidding No. 22/2023/QH15, the Government is the competent authority responsible for providing detailed regulations on the application of transitional provisions to business investment projects.<|im_end|>
```

B.2.2 Natural Language Inference

We formulated NLI as binary classification with task description, legal citation, and question. Models select "Yes" if the citation provides sufficient information to answer the question, or "No" otherwise.

```
<|im_start|>system
You are a Vietnamese legal expert with extensive experience in evaluating the relevance of legal documents to specific issues. Please provide an objective and accurate assessment.<|im_end|>
<|im_start|>user
Legal Text:
Penal Code 2017
Warning
A warning shall be applied to offenders who have committed less serious crimes and have multiple mitigating circumstances, but whose offenses do not yet warrant an exemption from punishment.

Question:
According to the Penal Code 2017, under what circumstances is a warning applied to an offender?

Does the above passage contain sufficient information to answer the question?
Please answer "Yes" or "No".<|im_end|>
<|im_start|>assistant
Yes<|im_end|>
```

B.2.3 Syllogism Reasoning

For Syllogism Reasoning, we employ step-by-step reasoning within `<think>` tags before generating the final answer in structured Major-Minor-Conclusion format.

```
<|im_start|>system
You are a Vietnamese legal expert. Please analyze and answer the legal question using a syllogistic reasoning structure:

Major Premise: Clearly state the applicable legal provisions (Laws, Decrees, Circulars, Articles).
Minor Premise: Analyze the specific situation and relevant factual elements.
Conclusion: Derive the legal consequences based on the application of the provisions to the specific situation.

Requirements: Accurately cite the relevant articles, and present logical and coherent arguments.<|im_end|>
<|im_start|>user
What fields and areas does the scope of regulation of the Law on Geology and Minerals No. 54/2024/QH15 cover?<|im_end|>
<|im_start|>assistant
<think>
Detailed Analysis: The Law on Geology and Minerals regulates activities related to the investigation, protection, extraction, processing, recovery, and management of geological and mineral resources. The scope of application of the Law covers not only inland areas but also extends to islands, internal waters, territorial seas, contiguous zones, exclusive economic zones, and the continental shelf. Furthermore, the Law also regulates activities related to financial management and the auctioning of mineral extraction rights. However, certain fields such as petroleum, natural water that is not natural mineral water, and certain mineral processing activities not associated with mineral extraction projects fall outside the regulated scope of the Law.
</think>

Major Premise: The Law regulates the scope of geology and minerals across the entire territory and maritime zones of Vietnam.
Minor Premise: Certain fields such as petroleum and natural water that is not natural mineral water fall outside the regulated scope.
Conclusion: The regulated scope of the Law encompasses activities related to the investigation, protection, extraction, processing, and management of geological and mineral resources within the inland areas, islands, maritime zones, and continental shelf of Vietnam, with the exception of fields expressly excluded under the provisions of the Law.<|im_end|>
```

B.3 Hyperparameter Settings

General Configuration All models fine-tuned on single NVIDIA P100 16GB GPU. Text sequences segmented using Pyvi’s ViTokenizer.⁵ We employ validation-based early stopping (patience 3) to prevent overfitting on synthetic training data.

⁵<https://github.com/trungtv/pyvi>

Encoder-based Models We use task-specific classification heads with discriminative fine-tuning. All models use AdamW optimizer, maximum sequence length 256 tokens, and weight decay 0.01. Table 9 presents task-specific configurations.

Table 9: Encoder fine-tuning hyperparameters.

Task	Learning rate	Batch size	Epochs
IR	1.5e-5	8	1
TF QA	2e-5	20	1
MCQ(-LK)	2e-5(3e-5)	20(12)	1(3)
EQA	2e-5	2	5
NLI	3e-5	16	5

IR: K=3 hard negatives.
EQA: sliding window stride 50.

Decoder-based Models We apply instruction tuning with QLoRA (Detmers et al., 2023): 4-bit NF4 quantization, LoRA rank 16, $\alpha=32$, dropout 0.05, adapters on all attention/FFN layers, FP16 training. Table 10 shows task configurations.

Table 10: Decoder fine-tuning hyperparameters.

Task	LR	BS	Grad	Acc	Epochs	Max Len	Gen
TF QA	3e-5*	4	1	1	1	2048	Logit [†]
MCQ(-LK)	2e-4*	4	4	1	1	2048	Logit [†]
AQA	4e-4*	2	4	1	1	4096	Greedy (4096)
NLI	2e-4*	4	2	1	1	2048	Logit [†]
Syllogism	4.5e-4*	1	1	1	1	4096	Greedy (2048)

*Paged AdamW 32-bit.

[†]Logit manipulation.

C LLM-as-a-Judge Criterion Breakdown for Syllogism Reasoning

Table 11 presents the per-criterion scores from LLM-Judge evaluation on the VLSP-Syllogism.

Table 11: Per-criterion LLM-as-a-Judge scores on VLSP-Syllogism. Each criterion contributes 0.25 points to the overall LLM-Judge score reported in Table 6.

Models	C1	C2	C3	C4
Qwen2-1.5B	0.1042	0.0139	0.1042	0.1319
Qwen2.5-1.5B	0.1319	0.0139	0.1319	0.1181
VLQ2.5-1.5B-Base (ours)	0.1181	0.0069	0.1181	0.1181
Qwen3-1.7B-Base	0.1597	0.0069	0.1597	0.0972
qwen3-1.7b-legal-pretrain	0.1667	0.0208	0.1667	0.1250
VLQ3-1.7B-Base (ours)	0.1667	0.0347	0.1667	0.1319
<i>gpt-4o-mini</i>	<i>0.4792</i>	<i>0.0833</i>	<i>0.4792</i>	<i>0.4653</i>

Each criterion is binary (0 or 1) per sample, and we report the mean score across 144 test samples. LLM scores each response on four binary criteria: whether the response is relevant to the question (C1), whether it includes accurate statutory references (C2), whether the legal reasoning is valid and grounded in the cited provisions (C3), and whether the conclusion correctly follows from the given rules and facts (C4).