
VietLegal

Benchmarking Vietnamese Legal Knowledge of Large Language Models

The first cognitively-grounded benchmark for Vietnamese civil-law LLM evaluation

Nguyen Tien Dong Minh-Anh Nguyen *et al.*

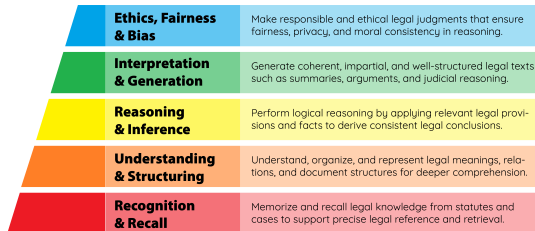
CMCAI · VinUniversity · HUST · SRH University Heidelberg



Motivation

- ▶ Legal NLP benchmarks are **English-centric & common-law**, built around case-based reasoning.
 - ▶ But **civil law governs >60% of jurisdictions**, authority comes from *codified statutes*, not precedent.
 - ▶ Civil-law challenges: deep **statutory hierarchies**, frequent **amendments**, **temporal validity**.
 - ▶ Vietnamese law adds intricate cross-references among **Articles / Clauses / Points** and **no standardized benchmark exists**.
- ⇒ VietLegal: first comprehensive Vietnamese legal benchmark, *replicable* to other civil-law jurisdictions.

Design of VietLegal: a five-level cognitive framework



The five-level cognitive framework of VietLegal

Bloom-inspired levels & metrics:

- ▶ **L1 Recognition & Recall** - Accuracy, EM (MCQ, NER)
- ▶ **L2 Understanding & Structuring** - Accuracy, Macro-F1, ROUGE-L (MCQ, MLC, BC, Gen.)
- ▶ **L3 Reasoning & Inference** - Accuracy, Binary-F1 (MCQ, BC)
- ▶ **L4 Interpretation & Generation** - ROUGE-L (Generation)
- ▶ **L5 Ethics, Fairness & Bias** - Accuracy (MCQ)

10,450 expert-annotated samples · 22 tasks · $\kappa = 0.89$

Experiment settings

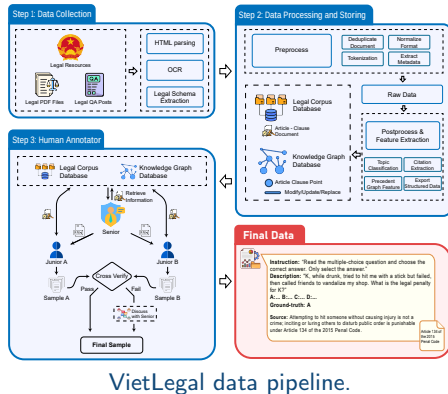
23 LLMs, two groups:

- **General multilingual** - GPT-4o / mini, Claude Sonnet 4.5, Gemini 2.5 Flash, Qwen2.5 (3B-72B), Llama-2/3.1, Gemma-2, InternLM
- **Vietnamese / domain-adapted** - SeaLLM-v3, BloomVN, Qwen3-4B-Legal, **CMC-Legal-32B**

Ablation designs:

- Zero-shot & few-shot × with / without CoT reasoning
- **Agentic RAG** over 55k statutes
- **Human evaluation** by legal experts

Temperature = 0 · middle-truncation for long inputs.



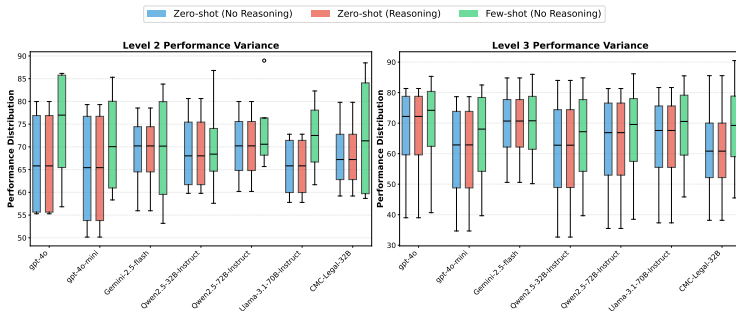
Main results - zero-shot across 22 tasks

- **Steep decline with cognitive complexity:** 80-90% at L1 recall → often <60% at L3 reasoning.
- **Task 3.1 (Article Prediction)** hardest - best overall ≈60% (Claude); best open-source 43.83%.
- **Task 1.5 (Schema Recall)** a universal bottleneck: *all models* <28%.
- **Conflict Detection (3.4):** degenerate behavior - "no-conflict" bias vs. over-detection.

Model Type	Model	Reasoning & Infer.					Interpret. & Generation			Ethics, Fairness & Bias			
		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
General Multilingual LLMs	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
	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
	CMC-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

Domain adaptation rivals scale: **CMC-Legal-32B** tops court-decision (90.67%) & unfair-contract (73.50%); proprietary lead shrinks at L3-L5.

Ablation I - Prompting strategies



- **Few-shot (>) zero-shot CoT:** demonstrations give *higher and more stable* accuracy across levels.
- **L2 "reasoning paradox":** explicit CoT often *hurts* understanding/structuring, holistic pattern-matching, not step-by-step.

- **L3 mixed & task-dependent:** CoT helps some procedural tasks (article ID) but stays unstable for conflict detection; few-shot reliably cuts variance.

⇒ Examples > forced reasoning; CoT is not uniformly helpful.

Ablation II - Human evaluation of generation tasks

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
CMC-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

Setup: double-blind expert rating, 1-5 Likert on **Legal Accuracy & Completeness**, Tasks 4.1-4.3, $\kappa = 0.92$.

- ▶ **Humans set the ceiling:** 4.76-4.88.
- ▶ Best model **CMC-Legal-32B**: 3.57-3.79 accuracy; GPT-4o second.
- ▶ **Persistent gap ≈ 1.2 -1.5 pts**, widest on Task 4.2 (IRAC reasoning).
 $\Rightarrow$ ROUGE-L alone is misleading; human judgment exposes the real gap.

Key takeaways

- ▶ **Advanced legal reasoning is largely unsolved** - article prediction & conflict detection remain hard.
- ▶ **Domain pretraining > scale**: small legal models beat models 10-18× larger.
- ▶ **Proprietary advantage shrinks** at higher cognitive levels.
- ▶ **Open problems for all models**: bias detection & legal-schema understanding.

We release **10,450 expert-verified samples** - a replicable standard for civil-law LLM evaluation.

Selected references I

-  Guha et al. (2023). *LegalBench: A Collaboratively Built Benchmark for Legal Reasoning*.
-  Chalkidis et al. (2022). *LexGLUE: A Benchmark Dataset for Legal Language Understanding*.
-  Li et al. (2024). *LexEval: A Comprehensive Chinese Legal Benchmark*.
-  Fei et al. (2024). *LawBench: Benchmarking Legal Knowledge of LLMs*.
-  Dai et al. (2025). *LAiW: A Chinese Legal LLM Benchmark*.
-  Merryman & Pérez-Perdomo (2018). *The Civil Law Tradition*.
-  Zheng et al. (2021). *CaseHOLD: When Does Pretraining Help?*
-  Wei et al. (2022). *Chain-of-Thought Prompting Elicits Reasoning in LLMs*.
-  Nguyen & Nguyen (2020). *PhoBERT: Pre-trained Language Models for Vietnamese*.
-  Phan et al. (2022). *ViT5: Pretrained Text-to-Text Transformer for Vietnamese*.