VerLexBench: A Version-Aware Benchmark for Measuring Temporal Drift in Legal QA

Anonymous ACL submission

Abstract

Current legal QA benchmarks grade models against today's version of the law, ignoring when the question was decided. Statutes change: in 71.1% of statute-citing Thai Supreme Court decisions, the cited law has changed since the court applied it. We introduce VerLexBench, a 1,367-item Thai legal QA benchmark across fourteen domains, anchoring every cited authority to its in-force text on the decision date. A controlled comparison swapping today's text for the in-force text on the same cited sections shows sixteen LLMs losing on both citation accuracy and reasoning quality, through two failure modes invisible to current evaluation: *confident-but-wrong-version*, where the model applies a section's current text instead of the historical one, and *missing-authority*, where the cited section has been repealed and is absent entirely. When the cited law has not changed since the decision, the gap disappears, confirming the cost is caused by the version mismatch. The static-snapshot assumption is a measurement choice rather than a domain constraint, and should be retired across longitudinal-corpus QA.

1 Introduction

Statutes change, but legal-NLP benchmarks do not track when they do. Existing Thai legal QA (NitiBench (Akarajaradwong et al., 2025)) and comparable benchmarks elsewhere (LawBench (Fei et al., 2024), LEXam (Fan et al., 2026), KCL (Oh et al., 2026), LegalBench-RAG (Pipitone and Alami, 2024), LegalAgentBench (Li et al., 2025)) evaluate retrieval against today's corpus, asking which sections to return but not which versions. An audit of the full Thai Supreme Court corpus finds that **in 71.1% of statute-citing decisions, at least one cited section has changed since the court applied it** (§3); a static-snapshot benchmark therefore scores systems against text the court did not apply on more than seven of every ten legally substantive cases.

The methodological gap is not Thai-specific. Any QA benchmark whose corpus is a longitudinal record of authoritative text carries the same confound: clinical guidelines, financial regulation, scientific standards, and software specifications all carry versioned provisions, yet no benchmark in any domain scores temporally correct retrieval over a versioned living document (Huwiler et al., 2025). The prescription, audit drift before reporting scores, applies across longitudinal-corpus NLP rather than only to legal text.

Why temporally-aware retrieval is load-bearing for legal practice. Compliance work routinely turns on temporally-correct authority: a tax audit asking what rate applied in 2020, or a labor dispute resting on rules in force at hire. A system that cites the latest version of an amended section gives a practitioner confident-but-wrong authority; a system that returns nothing on a now-repealed provision fails to find the law a 2008 court actually applied.

VerLexBench is a 1,367-item Thai legal QA benchmark across fourteen domains, with every cited authority anchored to its in-force law version on the question's decision date. The advisory track draws from rulings of eleven Thai government agencies (the agency's published reasoning is the gold answer); the case track draws from official decisions of five Thai courts (Supreme, Constitutional, Administrative, Tax, and Election), converted by an expert-free pipeline that structures judicial reasoning. Coverage extends Thai legal evaluation to intellectual property, environmental and natural resources, family and juvenile, immigration and nationality, and traffic and transport, areas underserved by prior legal NLP benchmarks.

Contributions of this work are as follows.

1. **Population-scale drift audit**: the first audit

at this scale on any legal corpus, finding that 71.1% of statute-citing Thai Supreme Court decisions cite text that has since changed (§3).

2. **Drift-cost measurement**: replacing in-force text with today's text costs every one of sixteen LLMs citation accuracy and reasoning quality across both perfect-retrieval (Gold-cited) and realistic-retrieval settings, through two named failure modes (§6).
3. **VerLexBench**: 1,367 items across fourteen Thai legal domains, every cited authority anchored to its in-force version at the decision date, scored on recall-weighted citation and the Thai Bar Council eight-tier reasoning rubric.

2 Related Work

Legal QA benchmarks. Existing legal QA benchmarks span English (LegalBench (Guha et al., 2023), LexGLUE (Chalkidis et al., 2022), CaseHOLD (Zheng et al., 2021)), Chinese (LawBench (Fei et al., 2024), LexEval (Li et al., 2024a)), and law-school exam reasoning (LEXam (Fan et al., 2026), KCL (Oh et al., 2026)); the only Thai benchmark, NitiBench (Akarajaradwong et al., 2025), is restricted to tax. Auxiliary work measures retrieval alone (LegalBench-RAG (Pipitone and Alami, 2024)), reports that retrieval gains transfer poorly to bar-exam QA (Zheng et al., 2025), or evaluates tool-using agents (LegalAgentBench (Li et al., 2025)). None of the anchors cited authority to the version in force on the decision date, so all of them implicitly score against a static snapshot of the law.

Legal NLP resources. Citation graphs and structural legal resources exist in several jurisdictions: AustLII's LawCite indexes Australasian case-law citations, LawNet-Viz (La Cava et al., 2022) models the Italian Civil Code as a citation network, and Lynx (Kaltenboeck et al., 2022) is a multilingual EU legal knowledge graph. None stores section text at multiple validity windows. FourCorners (Akarajaradwong et al., 2026b) is, to our knowledge, the first public legal resource that does, and VerLexBench uses it as a deterministic citation-resolution layer.

Temporal grounding. No prior legal benchmark scores retrieval against the in-force version on a decision date: LexTime (Barale et al., 2025) addresses event ordering within case text, and CLaw (Xu et al., 2025) timestamps Chinese statutes without posing date-conditioned tasks. The wider temporal-NLP literature addresses factoid freshness (Liška et al., 2022; Dhingra et al., 2022; Vu et al., 2024; Kasai et al., 2023) and event ordering (Chen et al., 2021; Zhang and Choi, 2021; Jang et al., 2022); the only versioned-document retrieval benchmark is VersionRAG (Huwiler et al., 2025), with 100 questions over software API changelogs. VersionRAG asks systems to retrieve a *named* version; VerLexBench fixes the version implicitly through the question's decision date and requires the system to discover the in-force snapshot itself, a strictly harder retrieval task over authoritative statutory text at thirteen times the item count.

LLM-as-judge for legal QA. Closed-API judges exhibit position and length biases and high inter-judge variance (Li et al., 2024b); LeMAJ (Enguehard et al., 2025) addresses this in legal QA via atomic Legal Data Point decomposition. VerLexBench adopts the Thai Bar Council eight-tier grading rubric directly, with judge selection grounded in an independent bar-exam reliability study reporting Krippendorff $\alpha = 0.77$ for a three-frontier-LLM panel against $\alpha = 0.36$ for a three-examiner human panel (Akarajaradwong et al., 2026a).

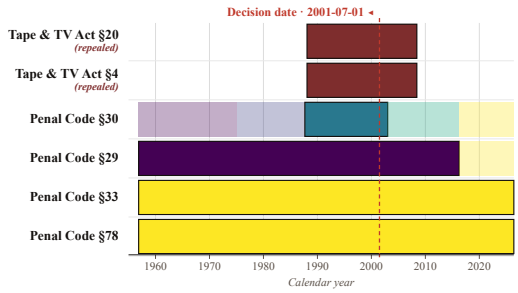

Figure 1: Six sections cited by Supreme Court Decision 9546/2544 (2001). Colored blocks: textual versions of the section; outlined block = version in force on the decision date (red dashed line). Oxblood = section in a since-repealed law. Only Penal Code section 33 and 78 still match today.

3 Measuring Drift in Thai Legal Practice

Measuring whether a court's cited text matches the active corpus requires a primitive that current legal-NLP datasets do not provide: the text of any cited section at any past date. We frame this as a *version-aware data layer* where each statute is represented

by a timeless abstract section and a sequence of content versions, each tagged with the date range during which it was in force. The query "what did the law say on date D" becomes a primary-key lookup, making drift between any cited text and the active corpus directly computable. Prior version-aware RAG (Huwiler et al., 2025) models version-state at retrieval time but does not anchor section text to its validity dates; without that anchor, a section-granular audit is not possible.

A cited section is *drifted* when its text has been amended between the decision date and today; Figure 1 illustrates this on Supreme Court Decision 9546/2544 (2001). We apply the data layer to the full Thai Supreme Court corpus (Akarajaradwong et al., 2026b).

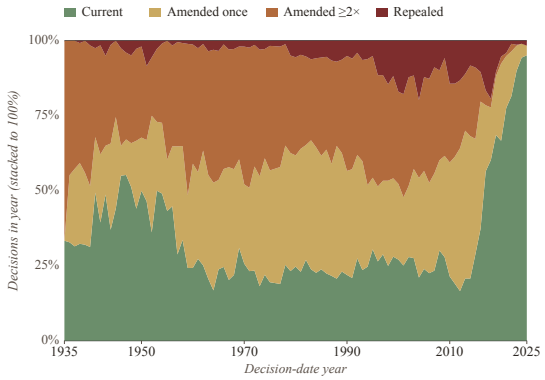

Figure 2: Decision-level drift severity by decision year (1935–2025), stacked to 100%. Each decision is classified by the most severe state of its citations: *current* (no drift), *amended once*, *amended* $\geq$2 times, or *repealed since the decision*. Overall: 29% / 32% / 31% / 8%; total drift 71.1%.

Across the Thai Supreme Court corpus (87,394 decisions), **in 71.1% of statute-citing decisions, the latest text of at least one cited section differs from the version the court applied**, either through textual amendment or because the law has since been repealed. Figure 2 decomposes this by decade and severity: each decision is classified by the most severe state of any of its citations into one of four bands. 29% are still *current* (no cited section has changed); 32% have at least one citation *amended once* since; 31% have at least one amended *two or more times*; and 8% cite at least one section whose parent law has been *repealed since the decision*.

The current band drops from around 40% in pre-1950s decisions to a 22% trough across the 1960s–1970s before recovering above 80% in the 2020s, while the repealed band grows from the 1980s onward to a 13% peak in the 2000s. Drift persists across decades: a static-corpus benchmark mismatches the deciding court text on more than seven of every ten legally substantive cases.

VerLexBench also retains a smaller class of items where the court cited a law that had already been *repealed* before the decision. Thai law requires courts to apply the law in force when the underlying act occurred, even if that law has since been repealed.[1] Supreme Court decisions 13798/2555, 6798/2544, and 6905/2544 illustrate the pattern, citing the Psychotropic Substances Act, the House of Representatives Election Act, and the House and Land Tax Act after each was repealed.

Because the cited text never changed since the court applied it, these items are not counted as drift; we report them in a separate *Retrospective* band. We keep them because they probe a regime no snapshot retriever can recover by construction: the cited authority has been removed from the active corpus entirely. Snapshot and temporal retrievers therefore fail symmetrically in operational settings, and only the Gold-cited factorial in §6 surfaces the resulting *missing-authority* cost. Historical-state retrieval for repealed laws is left to future work.

4 The VerLexBench Benchmark

VerLexBench contains 1,367 items across fourteen Thai legal domains and two tracks (advisory and case), drawn from authentic rulings of eleven Thai government agencies and decisions of five Thai courts (Supreme, Constitutional, Administrative, Tax, and Election). The advisory track uses agency rulings, which natively contain a factual scenario, legal question, authoritative answer, and cited provisions; the gold answer is the agency's published reasoning. The case track uses court decisions, which are not natively QA-shaped; an expert-free pipeline (§4.1) structures them into the same schema with the court's own reasoning as the gold answer. Both source classes are produced by legally trained authorities under official mandate and natively cite statutory authority by section number and statute name, providing deterministic citation ground truth without manual annotation.

Each domain caps at 100 items, so the equal-weighted macro-average tests multi-domain capability rather than tracking any one corpus's size;

[1] Penal Code §2–3 codify this for criminal matters; analogous intertemporal principles apply in other domains.

six domains include both tracks, eight are case-only (Thai legal practice provides no comparable advisory mechanism), and two cells fall short of the cap (securities and finance 97, family and juvenile 72), where the source pool is genuinely small. Per-domain source pools and final cell sizes are reported in Appendix A.

Citation anchoring and sub-questions. Every gold citation is anchored to its in-force version on the decision date via FourCorners (Akarajaradwong et al., 2026b), an externally maintained Thai legal resource that tracks section-level versions of every active Thai law; the released benchmark stores both the in-force text and the current-latest text, so resolved citations remain stable under future updates. Each item carries one or more sub-questions, each with its own verdict, reasoning, and load-bearing citation set; a dependency flag marks sub-questions whose verdict is conditioned on a prior one, and coupled chains are evaluated sequentially. Gold restricts to the minimum sufficient citation set the source authority marks as outcome-determinative for the sub-question's verdict; whole-law background mentions and source-narrative citations not relied on for the verdict are filtered at extraction and are not in the released data, so a model that retrieves them takes a precision hit even when the retrieval reflects thorough reading. Appendix B reproduces a full item with the released schema.

4.1 Extraction pipeline

The pipeline structures the source rather than rewriting it, admitting an item only when every cited authority resolves to a known section in FourCorners. Gemini 3.1 Pro Preview extracts a Pydantic record per source (fact pattern; sub-questions each with verdict, reasoning, citation list, and dependency flag; decision date; explicit source spans), classifying each citation as load-bearing or filterable. A GPT-5.4 verification pass gates faithfulness, dropping items whose extracted verdict, reasoning, or citation does not trace to a source span; survivors are stratified-sampled by decade, track, and sub-question count. Human annotators additionally reviewed a benchmark items against the source rulings to confirm extraction faithfulness. Gold reasoning and gold citations are taken verbatim from the source ruling, which was authored by the deciding court or issuing agency under official mandate, so VerLexBench inherits its legal correctness from the source authorities rather than from the annotator.

4.2 Temporal design

The corpus-scale drift documented in §3 carries through to VerLexBench at the sample scale: items are stratified across decades so each track contains a mix of stable and drifted citations. Each load-bearing citation is stored alongside both its in-force version on the decision date and the current latest version, yielding two corpora that differ only in amended sections: the **snapshot corpus**, the convention of prior legal QA benchmarks, and the **temporal corpus**, the content valid at each question's decision date. The two-corpus storage enables the Gold-cited factorial in §6, which isolates the version-mismatch cost from any retrieval error; snapshot retrieval mismatches the deciding authority on most items, while temporal-aware retrieval returns the version the court or agency actually applied.

5 Experimental Setup

5.1 Metrics

A VerLexBench answer is scored on two axes: *citation correctness* (deterministic) and *verdict-and-reasoning quality* (LLM-judge panel applying the official Thai Bar Council rubric).

Temporal Citation F$_\beta$. A load-bearing citation is a pair (s, h): the abstract section s and the content h of that section in force on the decision date. Given a sub-question's predicted citation set P and gold citation set G, the temporal-citation true-positive count $\mathrm{TP}^{\mathrm{TC}}$ is

$$\mathrm{TP}^{\mathrm{TC}} = \big|\{(p, g) \in P \times G : s_p = s_g \wedge h_p = h_g\}\big|,$$

and the headline metric is

$$\text{TC-F}_\beta = (1 + \beta^2) \cdot \frac{\mathrm{TP}^{\mathrm{TC}}}{\beta^2|G| + |P|},$$

yielding **TC-F2** ($\beta = 2$, recall-weighted, the headline because legal practice is recall-asymmetric) and **TC-F1** ($\beta = 1$, retained for continuity). A section-only variant **TC-F1$_{\text{naive}}$** substitutes $\mathrm{TP}^{\mathrm{naive}} = |\{(p, g) : s_p = s_g\}|$ in the same formula; this is the metric used by prior Thai legal QA benchmarks, and the TC-F1 minus TC-F1$_{\text{naive}}$ gap is the temporal-misalignment cost on that metric (tabulated in Appendix F). We also break out TC-precision and TC-recall to expose model-specific strategies (conservative high-precision emission versus broader exploration) that F$_\beta$ alone conflates.

	Reasoning quality				
Final answer	Unusable	Some	Acceptable	Good	Very good
matches gold	T4: 0–1	T5: 2–5	T6: 5–6	T7: 7–8	T8: 9–10
differs	T1: 0	T2: 1–2		T3: 2–4	

Table 1: Official Thai Bar Council eight-tier rubric, in force since 2014: grader picks a tier, then a half-integer score in its range.

Bar Council rubric. The Thai Bar Council eight-tier rubric (Table 1) crosses two correctness classes against five reasoning-quality levels and is the working standard for grading free-form Thai legal answers. VerLexBench adopts the rubric and the LLM-judge design from Akarajaradwong et al. (2026a): a three-frontier-LLM panel from three providers (Claude 4.6 Opus, Gemini 3.1 Pro Preview, GPT-5.4 Pro), scoring at temperature zero with the rubric reproduced verbatim and no chain of thought, tools, retrieval, or few-shot examples (Appendix D). That study reports panel reliability $\alpha = 0.77$ against human-examiner $\alpha = 0.36$ on the same 15 cross-graded answers.

Sub-question and domain aggregation. Both axes are computed per sub-question. For a metric m (TC-F$_\beta$ from above, or the panel-median rubric score across the three judges), the headline is the four-level macro-average

$$m = \frac{1}{|D|} \sum_{d \in D} \frac{1}{|I_d|} \sum_{i \in I_d} \frac{1}{|Q_i|} \sum_{q \in Q_i} m(q),$$

over domains D, items I_d per domain, and sub-questions Q_i per item. Macro-averaging at every level prevents items with many sub-questions, or domains with many items, from dominating the headline.

Reporting standard. TC-F2 is the primary deterministic headline; both TC-F2 and the Bar Council rubric are reported jointly in the main table (Table 2) and cost-quality Pareto figure (Figure 5), with per-domain and per-component breakdowns defaulting to TC-F2. $\beta = 2$ is chosen because the rubric tracks citation recall but not precision (§7).

5.2 Settings

VerLexBench evaluates systems on a factorial of four paradigms with a binary corpus axis (snapshot vs temporal); Parametric has no corpus split, the other three paradigms each have two variants, giving seven controlled settings on the shared items.

Parametric. Question and decision date only; no retrieval, no tools. Measures pretraining recall.

Gold-cited. Question, decision date, and the verbatim text of every gold-cited section. *Temporal*: each section fed at its in-force version on the decision date (reasoning ceiling under perfect retrieval). *Snapshot*: identical in every other respect, but each currently-active section fed at its latest version with sections whose parent law has since been repealed omitted, matching what a real snapshot retriever filtered on currently-active laws would return. The contrast isolates the cost of reading wrong-version text on correctly-identified sections.

Dense RAG. LLM agent with one tool, a Qwen3 dense-retrieval (Zhang et al., 2025) endpoint (selected per SEA-BED (Ponwitayarat et al., 2026) for Thai); top-k fixed at 10, with the agent choosing the query string and the number of queries. *Snapshot*: retriever indexes only the latest in-force version of each section and ignores the decision date; the agent's system prompt states the retrieval is detemporalised. *Temporal*: retriever indexes every dated version with validity windows, filtered server-side to the version covering the decision date.

FourCorners. LLM agent with read-only access to the FourCorners knowledge graph through Cypher queries against a public schema; DEKA-related node and edge types (court decisions and their citation edges) are not exposed. *Snapshot*: queries resolve to the latest in-force version of each section regardless of decision date, with the agent's system prompt stating this. *Temporal*: the schema exposes the version in force on the item's decision date.

Multi-agent harnesses and prompt-engineering interventions are deliberately excluded; the agent architecture is fixed across models, so performance differences reflect model capability.

5.3 Models and protocol

VerLexBench evaluates sixteen LLMs across six families: Claude Opus 4.7, Sonnet 4.6, Haiku 4.5; Gemini 3.1 Pro Preview, 3.5 Flash, 3 Flash Preview, 3.1 Flash Lite; GPT-5.4, -mini, -nano; DeepSeek V4 Pro and V4 Flash; Gemma 4 31B-IT and 26B-A4B-IT; Qwen 3.6 27B and 35B-A3B.

Zero-shot prompting; decoding at provider-default temperature. Per-(model, setting) TC-F1, TC-precision, TC-recall, and TC-F1$_{\text{naive}}$ are tabulated in Appendix F; cost-per-task is in Appendix G.

Model	Param.	Gold-cited		Agentic RAG		FC Agent	
		Snap.	Temp.	Snap.	Temp.	Snap.	Temp.
Claude Opus 4.7	**0.527 / 6.86**	**0.524 / 7.19**	0.884 / **7.39**	**0.392 / 6.91**	**0.628 / 7.06**	0.401 / **7.23**	**0.642 / 7.26**
Claude Sonnet 4.6	0.447 / 6.07	0.462 / 6.28	0.792 / 6.68	0.335 / 6.29	0.565 / 6.47	0.345 / 6.30	0.560 / 6.50
Claude Haiku 4.5	0.167 / 3.82	0.462 / 4.90	0.745 / 4.96	0.318 / 4.96	0.529 / 4.92	0.348 / 5.14	0.527 / 5.01
Gemini 3.1 Pro Preview	0.446 / 6.10	0.466 / 6.86	0.732 / 7.12	0.373 / 6.61	0.608 / 6.50	0.402 / 6.75	0.624 / 6.72
Gemini 3.5 Flash	0.464 / 6.37	0.496 / 7.13	0.845 / 7.15	0.391 / 6.87	**0.628** / 6.76	**0.404** / 6.76	0.628 / 6.86
Gemini 3 Flash Preview	0.444 / 5.94	0.499 / 6.78	0.859 / 6.80	0.346 / 6.12	0.562 / 6.25	0.391 / 6.26	0.619 / 6.65
Gemini 3.1 Flash Lite	0.367 / 5.17	0.448 / 5.55	0.782 / 6.38	0.289 / 5.15	0.471 / 5.38	0.324 / 5.24	0.531 / 5.74
GPT-5.4	0.417 / 6.24	0.515 / 6.82	**0.886** / 7.32	0.349 / 6.68	0.550 / 6.88	0.370 / 6.58	0.591 / 6.83
GPT-5.4-mini	0.380 / 5.14	0.504 / 5.95	0.839 / 5.81	0.321 / 5.23	0.533 / 5.45	0.327 / 5.46	0.500 / 5.56
GPT-5.4-nano	0.243 / 4.23	0.426 / 4.74	0.731 / 5.26	0.305 / 4.62	0.517 / 4.95	0.314 / 4.50	0.484 / 4.50
DeepSeek V4 Pro	0.306 / 5.29	0.413 / 5.67	0.637 / 6.19	0.365 / 6.20	0.593 / 6.42	0.388 / 6.52	0.606 / 6.46
DeepSeek V4 Flash	0.272 / 5.02	0.461 / 5.80	0.792 / 6.03	0.368 / 5.97	0.573 / 6.17	0.385 / 6.45	0.600 / 6.33
Gemma 4 31B-IT	0.336 / 5.38	0.420 / 6.15	0.617 / 6.28	0.282 / 6.05	0.494 / 6.20	0.333 / 6.08	0.549 / 6.28
Gemma 4 26B-A4B-IT	0.345 / 5.20	0.419 / 5.68	0.591 / 5.86	0.270 / 5.52	0.483 / 5.80	0.316 / 5.86	0.528 / 5.90
Qwen 3.6 27B	0.187 / 3.89	0.386 / 4.93	0.753 / 5.10	0.297 / 4.92	0.483 / 4.84	0.320 / 5.04	0.538 / 5.57
Qwen 3.6 35B-A3B	0.193 / 3.78	0.412 / 5.61	0.699 / 5.18	0.277 / 4.70	0.463 / 4.84	0.326 / 5.09	0.522 / 5.26

Table 2: Each cell shows TC-F2 / Bar Council rubric mean item score (0–10 scale) by model, retrieval paradigm, and temporal variant. Each retrieval paradigm appears in a snapshot and a temporally-aware variant; the snapshot-to-temporal contrast within a paradigm isolates the cost of evaluating against the wrong version of the cited law. Per-column maximum bolded and second-best underlined on each side of the slash independently. Frontier panel (top) above open-weight panel (below).

6 Results and Analysis

6.1 Overall performance

Table 2 reports TC-F2 and the Bar Council rubric mean item score by model and setting. Two panel-wide patterns anchor the analysis below. First, the *retrieval bottleneck*: Gold-cited columns reach TC-F2 0.59–0.89 across the panel while operational columns plateau at 0.46–0.64, exposing how much capability is lost between Gold-cited and realistic retrieval. Second, the *universality of the snapshot-to-temporal lift*: within every paradigm, every row in Table 2 shows a positive Δ on both metrics moving from the Snapshot to the Temporal column.

6.2 The temporal cost

Snapshot evaluation feeds the model the currently-latest version of each cited statute instead of the version in force on the decision date. The Gold-cited factorial isolates whether that text swap degrades reasoning when section identity is held fixed: both variants supply the correct cited sections (perfect retrieval), only the version differs. We partition items by the maximum-severity drift state of their load-bearing gold citations: *No drift* (control, cited text unchanged); *Amended* (text changed, parent law still active); *Repealed* (parent law no longer active); *Retrospective* (court cited an already-repealed law). Items can span multiple bands (Figure 1); per-item Δ averages over the item's stable sub-questions as well, so reported magnitudes are conservative.

Drift drives the cost. The no-drift band is the negative control: both axes register near-zero (Δ rubric -0.09, Δ TC-F2 -0.02), confirming that the factorial isolates the version swap rather than any prompt artifact. On the drifted bands the lift emerges sharply on both axes (Figure 3). The TC-F2 lift is partly mechanical: snapshot retrieval emits the latest version text by construction, so the text-match condition of TC-F2 fails regardless of model behavior. The rubric lift is the substantive cost. The rest of this section reads the cost on the rubric axis.

Two failure modes. The lift separates into two distinct modes. **Mode 1, confident-but-wrong-version** (44% of items, the Amended band): the model produces a defensible-looking verdict from the latest-version text but applies the section's current formulation rather than the historical formulation the court actually used. Rubric lift $+0.22$; TC-F2 lift $+0.53$. **Mode 2, missing-authority** (18% of items, the Repealed and Retrospective bands): the cited section is absent from the snapshot prompt, the model has no statutory text to reason from, and either hallucinates a substitute or refuses. Rubric lift $+0.85$ on Repealed, $+0.90$ on Retrospective (near a full tier on the 10-point scale); TC-F2 lifts $+0.48$ and $+0.31$. No existing benchmark holds

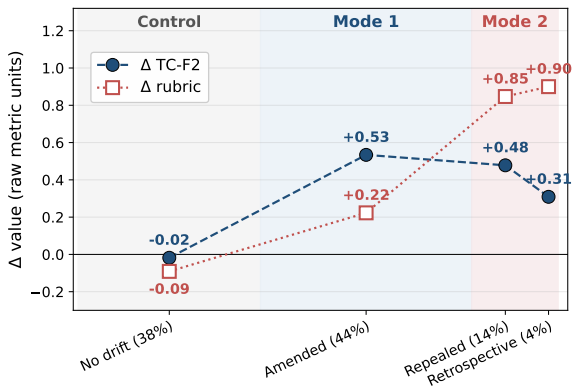

Figure 3: Gold-cited drift-band cost (Δ = Temporal − Snapshot), sixteen-model panel. TC-F2 (0–1 scale) and Bar Council rubric (0–10 scale) share the axis but are not directly comparable. Bands defined in §6.2; methodology in Appendix H.

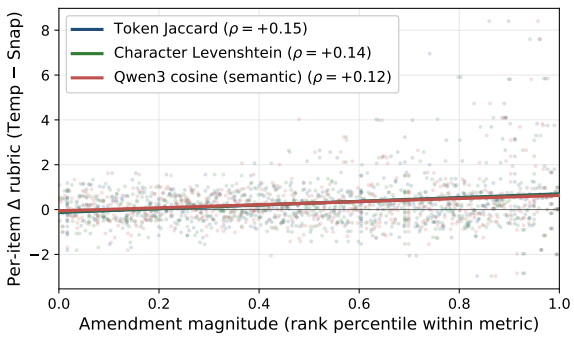

Figure 4: Amendment magnitude vs per-item Δ rubric across the 596 Mode 1 items. Each metric's X axis is rank-normalized so the OLS slope tracks the per-series Spearman ρ. Token Jaccard (blue), character Levenshtein (green), Qwen3 cosine (red).

the section fixed while varying the version, so both modes have been invisible to current legal-NLP evaluation.

Mode 1 cost is graded by amendment magnitude. Mode 1 items vary widely in how much the cited sections actually changed: some amendments are minor reformulations, others wholly repeal the provision. We measure amendment magnitude three ways per item: (1) 1− token Jaccard similarity (vocabulary turnover, Thai `newmm` tokenization (Phatthiyaphaibun et al., 2023)), (2) character-level Levenshtein distance (surface edits), and (3) 1− Qwen3 cosine similarity over the section's content embedding (semantic change). Across all Mode 1 items in VerLexBench, all three correlate positively with per-item rubric cost (Figure 4): Spearman $\rho = +0.15$ (Jaccard), $+0.14$ (Levenshtein), $+0.12$ (cosine), modest in magnitude but statistically robust ($p < 0.01$ across all three), consistent with amendment magnitude being one of several drivers of per-item cost rather than the dominant one. Surface metrics outperform the semantic embedding because Thai legal amendments typically preserve the section's topic while substantively changing the legal effect. None of these magnitudes predict TC-F2 cost ($\rho < 0.10$): TC-F2 fires on the binary presence of drift, while the rubric responds to its degree.

The two modes contribute unequally to the panel cost. Despite covering less than half as many items, Mode 2 contributes more to the panel-mean rubric drop than Mode 1: $18\% \times 0.85 = 0.15$ vs $44\% \times 0.22 = 0.10$ rubric points. Per-item, missing-authority is roughly $4\times$ as damaging as confident-but-wrong-version (0.85 vs 0.22 rubric points per item). The two modes are not equally tractable: Mode 1 is solvable by temporal-aware retrieval over the active corpus, while Mode 2 requires retrieval over already-repealed laws, a historical-state capability no current legal-NLP system provides.

The effect persists under realistic retrieval. The same pattern reproduces in operational settings at a smaller magnitude: panel-mean rubric lifts $+0.13$ to $+0.14$, and TC-F2 lifts $+0.17$ to $+0.24$, holding on every model (Table 2); the per-domain breakdown is in Appendix I. The cost is not specific to the Gold-cited setup. Combined with the population audit of §3, snapshot evaluation mismeasures the dominant case class across the Thai Supreme Court corpus rather than a marginal slice.

7 Discussion

Over-citation and metric choice. The metric defense rests on an empirical question: Does the Bar Council rubric, in fact, track recall over precision, justifying TC-F2's recall weighting? Within-setting Spearman correlations against the rubric reveal an asymmetry (Table 3): citation recall predicts the rubric strongly, while citation precision does not. The asymmetry is sharpest under FC, where over-citation is heaviest. TC-F2 inherits the recall-weighting and tracks the rubric across operational settings (per-setting Spearman 0.77 to 0.93); TC-F1's symmetric weighting attributes more of the system rank to a precision penalty that the rubric judges immaterial. Formalizing *sufficient-rather-than-exact citation* is left to future work.

Per-cell predictor	Spearman ρ	p
Precision	$+0.12$	0.12
Recall	$+0.71$	<0.001
TC-F1	$+0.62$	<0.001
TC-F2	$+0.77$	<0.001

Table 3: Mean within-setting Spearman correlation with the Bar Council rubric (sixteen models per setting, averaged across six settings). p from one-sample t-test on the six per-setting ρ values against zero ($df = 5$).

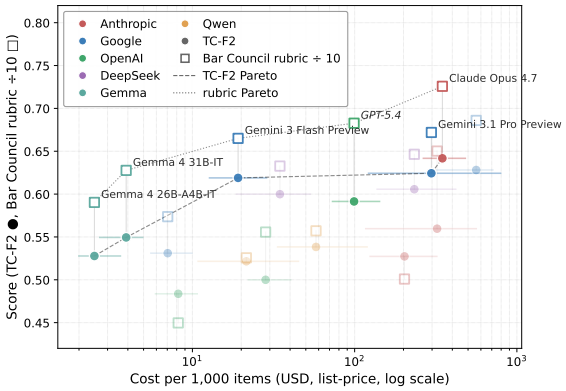

Figure 5: Cost-quality Pareto on Temporal FC. X: USD per 1,000 items (list price). Y: TC-F2 (filled circle), rubric/10 (hollow square). Whiskers: per-model cost IQR. Rate-card details in Appendix G.

Prior legal citation metrics have been overreporting Snapshot scores. Prior Thai legal QA benchmarks score citations on section identity alone and are structurally blind to which version of the section was retrieved. Table 4 shows the gap on our sixteen-model panel: section-only scoring inflates Snapshot systems by $+0.148$ to $+0.175$ TC-F1, mismeasuring them as 14 to 18 points more competent than the version-aware metric reveals.

Setting	TC-F1	TC-F1$_{\text{naive}}$	Inflation
Snapshot RAG	0.318	0.466	$+0.148$
Temporal RAG	0.492	0.492	0.000
Snapshot FC	0.344	0.519	$+0.175$
Temporal FC	0.547	0.547	0.000

Table 4: Panel-mean TC-F1 (version-aware, ours) vs TC-F1$_{\text{naive}}$ (section-only, prior-benchmark metric) on the operational settings. The inflation appears only on snapshot variants; temporal variants align by construction.

Cost-quality Pareto on Temporal FC. On Temporal FC, the highest-scoring operational cell across our settings, mid-tier and open-weight models cluster close to the frontier: Gemini 3 Flash Preview captures roughly 90% of Claude Opus 4.7's rubric at a 25× cost advantage (Figure 5). The rubric sits in Tier 7–8 ("good") across the panel, with no model reaching "very good" even at the top of the operational cost range.

Historical-state retrieval is the next benchmark target. Mode 2 is the more damaging failure mode per item (roughly 4× Mode 1) and contributes the largest share of the panel-mean cost (§6), yet no existing retrieval architecture addresses it. Snapshot retrievers filter out repealed laws by construction; temporal-aware retrieval over the active corpus solves Mode 1 but leaves Mode 2 intact. The missing capability is retrieval of already-repealed text, indexed by validity window and gated by the question's decision date, so the deciding-day text remains recoverable after the law is removed from the active corpus. Building this index is a data-layer rather than a modeling problem, consistent with the Gold-cited evidence that temporal correctness sits upstream of the LLM.

8 Conclusion

Legal-NLP evaluation has not accounted for the gap between when a court applies a statute and when a benchmark scores a model on it. VerLexBench closes the gap by anchoring every citation to the in-force text on the question's decision date, with a Gold-cited factorial that isolates the cost of reading the wrong version.

Drift is not marginal. Across the Thai Supreme Court corpus, the majority of statute-citing decisions reference at least one section that has drifted since the court applied it: the median condition over a versioned legal corpus, not an edge case.

Every model in our panel loses substantial citation accuracy and rubric quality on the snapshot ground truth, through two failure modes invisible to section-only metrics: *confident-but-wrong-version* (the model applies a section's current text rather than the historical one) and *missing-authority* (the cited section is absent from the snapshot).

Legal-NLP evaluation must treat drift as a first-class confound: audit it before reporting scores and locate temporal correctness in the data layer, not the model. This applies to any versioned corpus QA, from clinical guidelines to financial regulation.

Limitations

VerLexBench is positioned as a Thai-legal benchmark. The case-derived pipeline and the eight-tier rubric are designed around Thai legal practice;

cross-jurisdictional generalization is a natural future direction that we do not claim here.

Replicating the methodology in another jurisdiction requires a version-aware legal data layer of similar coverage to FourCorners (Akarajaradwong et al., 2026b), namely, section-level statutory text anchored to validity windows rather than a citation graph or amendment-metadata layer alone. To our knowledge, no comparable resource is currently public outside Thai law; building one is the natural prerequisite for cross-jurisdictional extensions.

The reasoning-quality axis uses a three-frontier-LLM judge panel applying the Thai Bar Council eight-tier rubric, with judge selection grounded in a bar-examination reliability study reporting Krippendorff $\alpha = 0.77$ for this panel (Akarajaradwong et al., 2026a). Reliability is anchored on the bar-exam study's coverage; examinees or domains substantially outside that coverage may warrant targeted validation.

Domain coverage reflects the structure of Thai legal practice rather than uniform sampling: criminal, constitutional, and electoral domains are case-only because no Thai agency issues advisory rulings on those subjects, and the Civil and Commercial advisory pool is small (15 items). We report all available advisory items in small-pool domains rather than padding with synthetic data.

The Gold-cited factorial swaps in-force text for current text while holding the model fixed. Because LLM pretraining is biased toward current statutory text, the snapshot prompt is partly aligned with the model's parametric prior, which could, in principle, inflate snapshot performance rather than cost it. The no-drift control band registers near-zero gap on both axes (§6.2), bounding this confound: under textually identical prompts, parametric alignment cannot explain a differential. A finer disentanglement of drifted items, isolating residual parametric contributions from in-context reasoning, is left to future work.

Ethics Statement

All rulings and decisions used are publicly available government documents. We remove personally identifiable information from inquiry texts during preprocessing and follow the agencies' and courts' published terms of use. VerLexBench is intended for research evaluation and is not a substitute for professional legal advice.

References

Pawitsapak Akarajaradwong, Wuttikrai Lertprasertphakorn, Chompakorn Chaksangchaichot, and Sarana Nutanong. 2026a. A two-phase stability study of llm judges and bar council examiners on thai bar-exam free-form essays. *Preprint*, arXiv:2605.25652.

Pawitsapak Akarajaradwong, Sarana Nutanong, and Chompakorn Chaksangchaichot. 2026b. Fourcorners: A production knowledge graph unifying thailand's legal system. In *The 64th Annual Meeting of the Association for Computational Linguistics – Industry Track*.

Pawitsapak Akarajaradwong, Pirat Pothavorn, Chompakorn Chaksangchaichot, Panuthep Tasawong, Thitiwat Nopparatbundit, Keerakiat Pratai, and Sarana Nutanong. 2025. NitiBench: Benchmarking LLM frameworks on Thai legal question answering capabilities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34304–34327, Suzhou, China. Association for Computational Linguistics.

Claire Barale, Leslie Barrett, Vikram Sunil Bajaj, and Michael Rovatsos. 2025. LexTime: A benchmark for temporal ordering of legal events. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5220–5236, Suzhou, China. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel 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, Dublin, Ireland. Association for Computational Linguistics.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Joseph Enguehard, Morgane Van Ermengem, Kate Atkinson, Sujeong Cha, Arijit Ghosh Chowdhury, Prashanth Kallur Ramaswamy, Jeremy Roghair, Hannah R Marlowe, Carina Suzana Negreanu, Kitty Boxall, and Diana Mincu. 2025. LeMAJ (legal LLM-as-a-judge): Bridging legal reasoning and LLM evaluation. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 318–337, Suzhou, China. Association for Computational Linguistics.

Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer,

Daniel Brunner, Markus Leippold, Mrinmaya Sachan, Alexander Stremitzer, Christoph Engel, Elliott Ash, and Joel Niklaus. 2026. Lexam: Benchmarking legal reasoning on 340 law exams. *Preprint*, arXiv:2505.12864.

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.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc.

Daniel Huwiler, Kurt Stockinger, and Jonathan Fürst. 2025. Versionrag: Version-aware retrieval-augmented generation for evolving documents. *Preprint*, arXiv:2510.08109.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Martin Kaltenboeck, Pascual Boil, Pieter Verhoeven, Christian Sageder, Elena Montiel-Ponsoda, and Pablo Calleja-Ibáñez. 2022. *Using a Legal Knowledge Graph for Multilingual Compliance Services in Labor Law, Contract Management, and Geothermal Energy*, pages 253–271. Springer International Publishing, Cham.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime qa: what's the answer right now? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lucio La Cava, Andrea Simeri, and Andrea Tagarelli. 2022. Lawnet-viz: A web-based system to visually explore networks of law article references. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3300–3305, New York, NY, USA. Association for Computing Machinery.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. 2025. LegalAgentBench: Evaluating LLM agents in legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2322–2344, Vienna, Austria. Association for Computational Linguistics.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024a. Lexeval: a comprehensive chinese legal benchmark for evaluating large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Adam Liška, Tomáš Kočiský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. *Preprint*, arXiv:2205.11388.

Hongseok Oh, Wonseok Hwang, and Kyoung-Woon On. 2026. Korean canonical legal benchmark: Toward knowledge-independent evaluation of LLMs' legal reasoning capabilities. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 232–243, Rabat, Morocco. Association for Computational Linguistics.

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntorntip, and Can Udomcharoenchaikit. 2023. PyThaiNLP: Thai natural language processing in Python. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 25–36, Singapore, Singapore. Empirical Methods in Natural Language Processing.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *Preprint*, arXiv:2408.10343.

Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Raymond Ng, Jann Railey Montalan, Thura Aung, Jian Gang Ngui, Yosephine Susanto, William Chandra Tjhi, Panuthep Tasawong, Erik Cambria, Ekapol

Chuangsuwanich, and Sarana Nutanong. 2026. Seabed: How do embedding models represent southeast asian languages? *Preprint*, arXiv:2508.12243.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. Fresh-LLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Xinzhe Xu, Liang Zhao, Hongshen Xu, and Chenchenc. 2025. CLaw: Benchmarking Chinese legal knowledge in large language models - a fine-grained corpus and reasoning analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12071–12103, Suzhou, China. Association for Computational Linguistics.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

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 of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 159–168, New York, NY, USA. Association for Computing Machinery.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. 2025. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, CSLAW '25, page 169–193, New York, NY, USA. Association for Computing Machinery.

A Source Inventory

Table 5 reports the per-domain source pool and final per-cell sizes across both tracks. Advisory sources: **RD** (Revenue Department), **Excise** (Excise Department), **SEC** (Securities and Exchange Commission), **AMLO** (Anti-Money Laundering Office), **DBD** (Department of Business Development), **DOPA** (Department of Provincial Administration), **OCS** (Office of the Council of State), **Anamai** (Department of Health), **OCPB** (Office of the Consumer Protection Board), **FDA** (Food and Drug Administration), **DIT** (Department of Internal Trade). Case sources: **DEKA** (Supreme Court), **ConC** (Constitutional Court), **AdmC** (Administrative Court), **TaxC** (Tax Court), **ECT** (Election Commission of Thailand). Raw-pool counts are the number of source rulings or decisions ingested from each authority before extraction-faithfulness filtering and per-cell capping; final cell sizes reflect the cap of 100 items per (domain, track) pair, except securities and finance (97) and family and juvenile (72) where the source pool was genuinely smaller after filtering.

B Released Schema and Dataset Example

Each item in the released dataset is JSON with the schema below; the worked example populates every field on a real Securities and Exchange Commission advisory ruling (decision date 2001-06-19).

Item. `id`, `domain` (one of fourteen Thai legal domains), `track` (ADVISORY or CASE), `source` (agency or court short tag), `source_id`, `decision_date` (ISO-8601), `year_be`, `decade`, `title`, `source_text` (full ruling text, for closed-book and RAG use), `fact_pattern`, `sub_questions`.

Sub-question. `question`, `verdict`, `reasoning`, `cited_laws` (grouped by parent law), `depends_on_previous` (boolean; TRUE when this sub-question's verdict is conditioned on the prior one).

Cited law. `law_name` (as cited in the source), `law_code` (FourCorners canonical identifier), `law_title` (canonical full title), `components` (list of cited sections within this law).

Component. `label` (section label as cited), `abstract_uri` (`th/law/{law_code}/section-{label}`, the timeless identifier), and a pair of version blocks. The in-force block (`version_uri`, `valid_from`, `valid_to`, `content_hash`, `content`) carries the version that was in force on `decision_date`. The latest-snapshot block (`latest_version_uri`, `latest_valid_from`, `latest_content_hash`, `latest_content`) carries the current-latest in-force version of the same section, so consumers can build Gold-cited Snapshot prompts with no

	Advisory track		Case track		
Domain	**Sources (raw pool)**	**Items**	**Sources (raw pool)**	**Items**	**Total**
Tax & Revenue	RD (4,658), Excise (30)	47	TaxC (73), DEKA†	52	99
Securities & Finance	SEC (391), AMLO (16)	28	DEKA†	69	97
Civil & Commercial	DBD (15)	2	DEKA†	98	100
Criminal	none	0	DEKA†	99	99
Administrative	DOPA (124), OCS (14,897)	9	AdmC (13,093), DEKA†	91	100
Labor	none	0	DEKA†	100	100
Constitutional & Election	none	0	ConC (1,685), ECT (5,376), DEKA†	100	100
Consumer & Health	Anamai (363), OCPB (106), FDA (15), DIT (22)	30	DEKA†	70	100
Land & Property	none	0	DEKA†	100	100
Intellectual Property	none	0	DEKA†	100	100
Environment & Nat. Resources	none	0	DEKA†	100	100
Family & Juvenile	none	0	DEKA†	72	72
Immigration & Nationality	none	0	DEKA†	100	100
Traffic & Transport	none	0	DEKA†	100	100
Total	*11 agencies*	**116**	*5 courts*	**1,251**	**1,367**

Table 5: Source inventory and final per-cell sizes. †DEKA: Thai Supreme Court corpus, assigned to a domain via the FourCorners citation network (Akarajaradwong et al., 2026b).

FourCorners dependency; all four fields are `null` when the section's parent law has since been repealed. Convenience flags `is_drifted` (TRUE when `content_hash` $\neq$ `latest_content_hash`) and `is_repealed` (TRUE when the parent law is no longer active) let consumers separate Mode 1 from Mode 2 in one branch.

All components in the released gold are load-bearing; whole-law background mentions and citations not relied on for the verdict are filtered at publish time.

> **Item** `sec_6889`. *Domain*: securities and finance. *Track*: advisory. *Decision date*: 2001-06-19.
>
> **Fact pattern** (Thai, translated): A mutual fund leasing real property must provide security, cleaning, and maintenance services. The fund management company hires a property manager that subcontracts these services. The property manager fears that if the mutual fund cancels its engagement, it will be left owing unpaid sums to subcontractors and seeks a bank guarantee from a commercial bank.
>
> **Sub-question 1**: May a mutual fund have a commercial bank issue a bank guarantee for the benefit of the property manager? Does this create an encumbrance prohibited by section 126(5) of the Securities and Exchange Act B.E. 2535?
>
> **Gold verdict**: The proposed arrangement is *not* prohibited under section 126(5). The general phrase "create any encumbrance" must be read narrowly so as not to capture ordinary commercial activity (e.g., installment-purchase obligations, contractual penalty clauses).
>
> **Gold citation**: Securities and Exchange Act B.E. 2535, section 126. *abstract_uri*: `th/law/L0015-1B-0002/section-126`. *In-force on decision date*: `version_uri` `.../section-126/v/2000-03-30`, valid 2000-03-30 to 2003-07-10, `content_hash` `7449540f5756`. *Latest snapshot*: `latest_content_hash` `0cfc19385cb6`, `is_drifted` TRUE, `is_repealed` FALSE. Table 6 traces the section's full version history; a snapshot retriever would return the latest text and miss the version the court actually applied.

The data release includes full Thai source text and additional examples for each domain. Table 6 traces the four distinct textual versions of section 126 over time, illustrating the drift pattern that the version-aware data layer resolves.

content_hash	valid_from	valid_to
`9a9450ba6bab`	1992-05-15	2000-03-29
`7449540f5756`	2000-03-30	2003-07-10
`973ae57705e5`	2003-07-11	2016-12-11
`0cfc19385cb6`	2016-12-12	present

Table 6: Distinct textual versions of Securities and Exchange Act §126 (`th/law/L0015-1B-0002/section-126`). The 2001 SEC ruling above applies hash `7449540f5756` in force at the decision date; a snapshot retriever would return hash `0cfc19385cb6`, the current latest. Three subsequent amendments separate the two.

C Bar Council Eight-tier Rubric

The full grading regulation reproduced inline in the LLM-judge prompt is shown in Table 1 of the body. The regulation was issued by the Office of Legal Education Training of the Thai Bar Council in 2014 and remains the standardised grading instrument for Thai legal essay answers. We reproduce here a faithful English translation of the scoring clause for completeness; the verbatim Thai text is included in the data release.

The grading considers the final answer and supporting reasons together, on a per-question basis:

(1) Final answer differs from gold, reasoning unusable: 0.
(2) Final answer differs, reasoning somewhat acceptable: 1 to 2.
(3) Final answer differs, reasoning good: 2 to 4.
(4) Final answer matches gold, reasoning unusable: 0 to 1.
(5) Final answer matches gold, reasoning somewhat acceptable: 2 to 5.
(6) Final answer matches gold, reasoning acceptable: 5 to 6.
(7) Final answer matches gold, reasoning good: 7 to 8.
(8) Final answer matches gold, reasoning very good: 9 to 10.

If a question contains multiple issues, the grader fixes the score for each issue first and then applies the criteria to each. The score ranges overlap (tier 3 caps at 4, above tier 4's cap at 1), so the numeric score alone does not recover the tier; we report both tier and score per item. For multi-sub-question advisory items, we apply the rubric per sub-question and average within an item; the modal tier is reported, ties broken toward the lower tier.

D Judge Prompt

The prompt below is identical across judge models and across all items; only the question text, decision date, gold answer, and model answer change per item. Sampling: temperature 0, fixed seed where available, single forward pass, no chain-of-thought scaffolding, no tool use, no retrieval, no few-shot examples.

D.1 System prompt (English translation)

```
You are an expert grader of Thai legal answers. You will apply the official Thai Bar Council grading regulation (reproduced below) to the given answer. The regulation is the same instrument used by trained Thai bar examiners.

You have four inputs: (1) the legal question, (2) the decision date on which the question was answered (used to determine which version of cited statutes was in force), (3) the official gold answer, and (4) the model answer to be graded.

You must produce:

  • A tier number (1 to 8) from the regulation that best fits the model answer.
  • A numeric score within that tier's score range (half-integers allowed).
  • A one-paragraph justification (Thai) citing the decisive statutory sections and explaining why the answer falls into the chosen tier.

For items with multiple sub-questions, fix the tier and score for each sub-question independently, then return a list of {tier, score, justification} triples in sub-question order.

Return strict JSON. Do not include any prose outside the JSON.

Thai Bar Council Grading Regulation: see Appendix C for the verbatim text reproduced inline in the prompt.
```

D.2 User message schema

```
QUESTION: {question text}
DECISION DATE: {ISO-8601 date, e.g. 2018-03-15}
OFFICIAL GOLD ANSWER: {gold answer extracted from source ruling or decision}
MODEL ANSWER: {model answer to be graded}
```

D.3 Expected JSON output (single sub-question)

```
{ "tier": 7, "score": 7.5, "justification": "[one-paragraph Thai legal justification citing relevant statutory sections]" }
```

D.4 Expected JSON output (multi sub-question advisory item)

```
{ "sub_questions": [
  { "tier": 7, "score": 7.5, "justification": "..." },
  { "tier": 5, "score": 4, "justification": "..." },
  { "tier": 8, "score": 9, "justification": "..." }
] }
```

Inputs the judge does not receive. The judge receives no information about the model that produced the answer, no information about the system or setting under which the answer was produced (Closed-Book, RAG-Snapshot, etc.), no scores from any other judge, and no access to the Four-Corners graph. TC-F1 (the deterministic citation axis) is computed independently of the judge's reasoning score. The judge sees only the four inputs listed above.

E Metric Implementation Details

Temporal Citation F_β is implemented in two stages. Resolution maps a model's emitted citation $(\text{law_name}, \text{section_label})$ or canonical URI to an abstract section identifier and, given a date, to the in-force version's text content.

1. **Citation parsing.** The benchmark's emitted citation schema carries optional URI, law_name, and section fields. When a model emits a canonical `th/law/.../section-N` URI it is taken verbatim; otherwise the law_name is resolved against the law catalog via `rapidfuzz.WRatio` (threshold ≥ 85) and the section label is normalised (Thai numerals to Arabic, sub-paragraph parens stripped: *section-3 (5)* $\rightarrow$ *section-3*) before lookup. WRatio was chosen over plain Levenshtein after year-suffixed law names (e.g., "Labour Protection Act" vs "Labour Protection Act B.E. 2541") failed the threshold under plain ratio.
2. **Version anchoring.** Given an abstract section and a lookup date, a single Cypher query against FourCorners returns the in-force ComponentVersion whose `PART_OF` edge has `valid_from` $\leq$ date $\leq$ `valid_to`, together with that version's `content_hash`. For snapshot settings the lookup date is fixed to `9999-12-31` so the latest version is returned and the predicted hash reflects the text the agent actually read; for temporal settings the lookup uses the item's decision date so a correct section identification matches the gold hash.
3. **Aggregation.** Precision, recall, and F_β are computed per sub-question over the set of (s, h) pairs, averaged within an item, macro-averaged within a domain, then macro-averaged across the fourteen domains.

Edge cases. An item with empty gold and empty predicted citations scores vacuously 1.0. A predicted URI that resolves to a real abstract section but has no in-force version on the lookup date records hash $= \varnothing$, which never matches gold under m^{TC} but can match under m^{naive}. Malformed URIs (`https://example.com/...`, pseudo-labels such as `Narcotics_Act_2564_Section_90`) are treated as missing and fall through to the fuzzy resolver.

E.1 Worked Scoring Examples

Three worked examples illustrate m^{TC} vs m^{naive} on a sub-question with gold load-bearing set $G = \{(s_1, h_1^*), (s_2, h_2^*)\}$, where h_i^* denotes the in-force text content of section s_i at the decision date and $h_i^\dagger$ denotes the current latest text (latest $\neq$ in-force, i.e., the section drifted).

1. *Exact temporal match.* Predicted $P = \{(s_1, h_1^*), (s_2, h_2^*)\}$. Then m^{TC} matches both; precision $=$ recall $= F_1 = F_2 = 1$.
2. *Right section, wrong version (snapshot retrieval against drift).* Predicted $P = \{(s_1, h_1^\dagger), (s_2, h_2^\dagger)\}$. m^{TC} matches zero; $F_1 = F_2 = 0$. m^{naive} matches both on section identity alone; $F_1^{\text{naive}} = F_2^{\text{naive}} = 1$. The gap $F_1^{\text{naive}} - F_1 = 1$ is the temporal-misalignment cost on this item.
3. *Partial precision with one drifted citation.* Predicted $P = \{(s_1, h_1^*), (s_2, h_2^\dagger), (s_3, h_3^*)\}$ where $s_3 \notin G$. m^{TC} matches only (s_1, h_1^*): TP $= 1$, $|P| = 3$, $|G| = 2$, so precision $= 1/3$, recall $= 1/2$, $F_1 = 0.40$, $F_2 = 0.45$. m^{naive} matches $\{(s_1), (s_2)\}$ but ignores hashes: precision $= 2/3$, recall $= 1$, $F_1^{\text{naive}} = 0.80$, $F_2^{\text{naive}} = 0.91$.

F Per-Model Results

The primary TC-F2 and Bar Council rubric panel (Table 2) is reported in the main experimental setup section. This appendix supplies the per-component breakdown (TC-precision, TC-recall), TC-F1 (the harmonic-mean baseline retained for continuity with prior Thai legal QA work), and the section-only Citation F1 ablation (TC-F1$_{\text{naive}}$, the metric used by prior benchmarks). Per-component values surface model-specific emission strategies: a high TC-precision with a moderate TC-recall indicates conservative "cite only what you are sure of" behaviour, whereas a moderate TC-precision with a high TC-recall indicates broader exploration. All four tables share the body main table's paradigm-grouped layout and are rendered from per-cell scorer output by the same script.

G Cost Accounting and Rate Cards

Every cost number reported in this paper is recomputed from a single per-model rate card (Table 11) applied to the realized prompt and completion token counts of each run row. No prompt-caching, flex-tier, or batch-API discount is folded in: the same standard-tier list price is charged for every token of every model. Open-weight self-hosted models (Gemma, Qwen) are priced at their published OpenRouter serving rate. This keeps the cost comparison reproducible from token counts alone and consistent across providers. Per-model per-setting medians are in Table 12.

Model	Param.	Gold-cited		Agentic RAG		FC Agent	
		Snap.	Temp.	Snap.	Temp.	Snap.	Temp.
Claude Opus 4.7	**0.509**	**0.538**	**0.899**	0.356	0.560	0.380	0.612
Claude Sonnet 4.6	0.436	0.483	0.819	0.286	0.481	0.295	0.488
Claude Haiku 4.5	0.168	0.489	0.779	0.299	0.493	0.328	0.506
Gemini 3.1 Pro Preview	0.470	0.492	0.758	0.381	**0.619**	**0.397**	**0.616**
Gemini 3.5 Flash	0.474	0.517	0.869	**0.382**	0.607	0.385	0.599
Gemini 3 Flash Preview	0.446	0.521	0.883	0.349	0.555	0.382	0.601
Gemini 3.1 Flash Lite	0.379	0.477	0.820	0.307	0.495	0.342	0.556
GPT-5.4	0.396	0.526	0.896	0.308	0.479	0.331	0.547
GPT-5.4-mini	0.370	0.524	0.865	0.312	0.511	0.315	0.490
GPT-5.4-nano	0.254	0.455	0.766	0.309	0.506	0.319	0.493
DeepSeek V4 Pro	0.317	0.440	0.672	0.342	0.531	0.361	0.568
DeepSeek V4 Flash	0.282	0.488	0.826	0.333	0.510	0.347	0.527
Gemma 4 31B-IT	0.362	0.454	0.659	0.304	0.521	0.351	0.569
Gemma 4 26B-A4B-IT	0.368	0.451	0.630	0.292	0.510	0.337	0.551
Qwen 3.6 27B	0.184	0.414	0.796	0.302	0.487	0.309	0.518
Qwen 3.6 35B-A3B	0.201	0.446	0.743	0.291	0.469	0.317	0.513

Table 7: TC-F1 (Temporal Citation F1) by model and setting. Per-column maximum bolded, second-best underlined. − = pending or completion rate below 80%. Frontier panel (top) above open-weight panel (below).

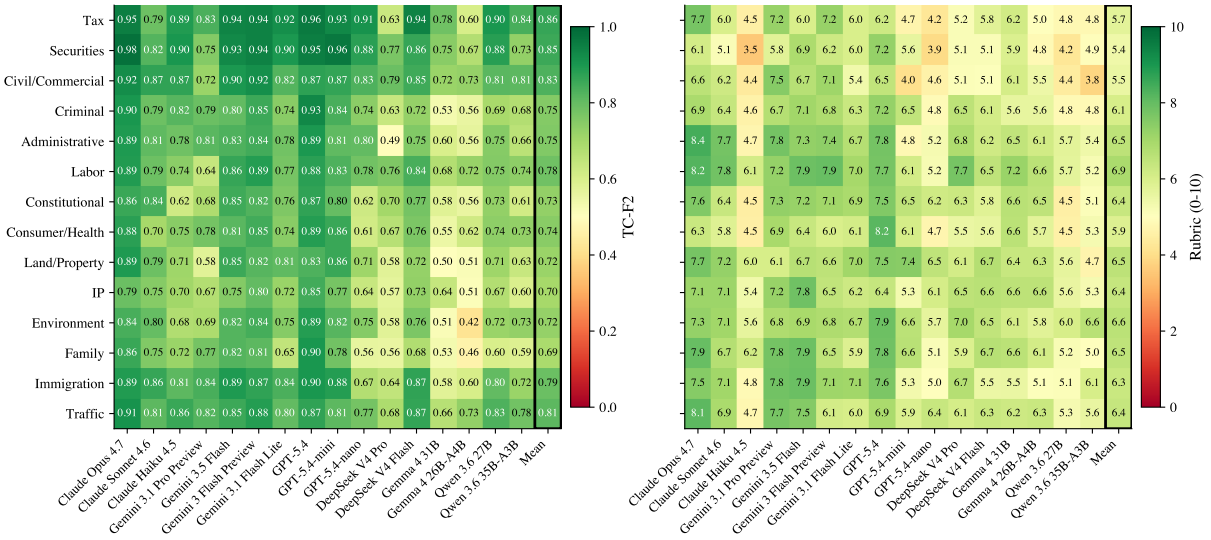

Figure 6: Per-(model, domain) reasoning capability on the Gold-cited setting (perfect retrieval): TC-F2 (left) and Bar Council rubric mean item score (right). Rightmost *Mean* column (separated by a blank gap) is the panel mean per domain.

H Drift-Band Methodology

Items are classified by the maximum-severity drift state across their load-bearing gold citations using the §3 audit Cypher; per-item Δ averages over the item's citations and may include unchanged citations that dilute the band-level magnitude. Panel-mean Δ per band is the unweighted mean over (model, item) pairs ($N \approx 1{,}720$ / 1,885 / 690 / 110 for No-drift / Amended / Repealed / Retrospective). On the drifted bands every one of sixteen models shows the same sign on both axes (panel sign-consistency: $16/16$ for TC-F2 on Amended and Repealed; $14/16$ for the rubric on Amended, $16/16$ on Repealed). The No-drift control's Δ TC-F2 $= -0.02$ and Δ rubric $= -0.09$ are indistinguishable from zero, confirming the cost on drifted bands traces to version mismatch rather than to any prompt artefact.

I Per-Domain Temporal Cost

Table 13 reports per-domain panel-mean TC-F2 lift under both retrieval paradigms; the per-(model, domain) heatmap (Figure 6) gives the matching reasoning-ceiling view under Gold-cited retrieval.

Model	Param.	Gold-cited		Agentic RAG		FC Agent	
		Snap.	Temp.	Snap.	Temp.	Snap.	Temp.
Claude Opus 4.7	0.549	**0.613**	0.963	0.360	0.534	0.400	0.640
Claude Sonnet 4.6	0.483	0.577	0.916	0.262	0.432	0.264	0.451
Claude Haiku 4.5	0.193	0.604	0.908	0.318	0.513	0.349	0.545
Gemini 3.1 Pro Preview	**0.586**	0.591	0.859	**0.459**	**0.723**	0.445	0.679
Gemini 3.5 Flash	0.560	0.609	0.961	0.427	0.649	0.407	0.624
Gemini 3 Flash Preview	0.513	0.607	**0.971**	0.426	0.627	0.427	0.651
Gemini 3.1 Flash Lite	0.473	0.599	0.967	0.421	0.635	**0.455**	**0.698**
GPT-5.4	0.415	0.587	0.947	0.301	0.446	0.320	0.546
GPT-5.4-mini	0.400	0.612	0.963	0.350	0.556	0.344	0.550
GPT-5.4-nano	0.321	0.590	0.920	0.378	0.573	0.385	0.592
DeepSeek V4 Pro	0.389	0.561	0.803	0.360	0.517	0.363	0.583
DeepSeek V4 Flash	0.336	0.607	0.962	0.327	0.493	0.333	0.494
Gemma 4 31B-IT	0.483	0.594	0.823	0.416	0.661	0.453	0.694
Gemma 4 26B-A4B-IT	0.486	0.578	0.772	0.409	0.658	0.442	0.679
Qwen 3.6 27B	0.212	0.536	0.963	0.367	0.580	0.332	0.554
Qwen 3.6 35B-A3B	0.252	0.592	0.918	0.372	0.556	0.353	0.579

Table 8: Precision on temporal citation pairs (s, h) by model and setting. Per-column maximum bolded, second-best underlined. $-$ = pending or completion rate below 80%. Frontier panel (top) above open-weight panel (below).

Model	Param.	Gold-cited		Agentic RAG		FC Agent	
		Snap.	Temp.	Snap.	Temp.	Snap.	Temp.
Claude Opus 4.7	**0.562**	**0.520**	0.880	**0.445**	**0.725**	0.435	0.691
Claude Sonnet 4.6	0.476	0.455	0.782	0.411	0.689	0.420	0.669
Claude Haiku 4.5	0.170	0.451	0.730	0.350	0.590	0.384	0.568
Gemini 3.1 Pro Preview	0.440	0.455	0.721	0.379	0.618	0.420	0.652
Gemini 3.5 Flash	0.469	0.488	0.836	0.415	0.670	0.437	0.679
Gemini 3 Flash Preview	0.456	0.490	0.849	0.359	0.587	0.413	0.656
Gemini 3.1 Flash Lite	0.369	0.436	0.765	0.288	0.467	0.321	0.526
GPT-5.4	0.451	0.513	**0.887**	0.411	0.653	0.428	0.660
GPT-5.4-mini	0.402	0.498	0.830	0.342	0.575	0.352	0.529
GPT-5.4-nano	0.244	0.414	0.716	0.316	0.551	0.322	0.493
DeepSeek V4 Pro	0.305	0.402	0.621	0.405	0.689	0.431	0.668
DeepSeek V4 Flash	0.270	0.449	0.777	0.425	0.672	**0.443**	**0.716**
Gemma 4 31B-IT	0.327	0.406	0.598	0.275	0.488	0.331	0.549
Gemma 4 26B-A4B-IT	0.337	0.405	0.573	0.263	0.477	0.310	0.524
Qwen 3.6 27B	0.196	0.374	0.733	0.304	0.500	0.346	0.579
Qwen 3.6 35B-A3B	0.193	0.396	0.679	0.274	0.476	0.347	0.551

Table 9: Recall on temporal citation pairs (s, h) by model and setting. Per-column maximum bolded, second-best underlined. $-$ = pending or completion rate below 80%. Frontier panel (top) above open-weight panel (below).

Model	Param.	Gold-cited		Agentic RAG		FC Agent	
		Snap.	Temp.	Snap.	Temp.	Snap.	Temp.
Claude Opus 4.7	**0.509**	**0.818**	**0.902**	0.532	0.560	0.583	<u>0.612</u>
Claude Sonnet 4.6	0.436	0.731	0.822	0.413	0.481	0.438	0.488
Claude Haiku 4.5	0.168	0.751	0.780	0.434	0.493	0.484	0.506
Gemini 3.1 Pro Preview	0.470	0.737	0.758	**0.564**	**0.619**	**0.605**	**0.616**
Gemini 3.5 Flash	<u>0.474</u>	0.785	0.872	<u>0.552</u>	<u>0.607</u>	<u>0.595</u>	0.599
Gemini 3 Flash Preview	0.446	0.783	0.885	0.506	0.555	0.583	0.601
Gemini 3.1 Flash Lite	0.379	0.737	0.822	0.450	0.495	0.525	0.556
GPT-5.4	0.396	<u>0.809</u>	<u>0.899</u>	0.454	0.479	0.493	0.547
GPT-5.4-mini	0.370	0.800	0.869	0.456	0.512	0.476	0.490
GPT-5.4-nano	0.254	0.695	0.768	0.454	0.506	0.473	0.493
DeepSeek V4 Pro	0.317	0.683	0.673	0.494	0.532	0.539	0.568
DeepSeek V4 Flash	0.282	0.750	0.827	0.492	0.510	0.516	0.527
Gemma 4 31B-IT	0.362	0.690	0.662	0.459	0.521	0.536	0.569
Gemma 4 26B-A4B-IT	0.368	0.688	0.631	0.435	0.510	0.517	0.551
Qwen 3.6 27B	0.184	0.643	0.798	0.450	0.487	0.476	0.518
Qwen 3.6 35B-A3B	0.202	0.697	0.745	0.413	0.469	0.465	0.513

Table 10: TC-F1$_{\text{naive}}$ (section-only ablation; matches the prior-benchmark metric) by model and setting. Per-column maximum bolded, second-best underlined. $-$ = pending or completion rate below 80%. Frontier panel (top) above open-weight panel (below).

Model	Input $/M	Output $/M
Claude Opus 4.7	5.00	25.00
Claude Sonnet 4.6	3.00	15.00
Claude Haiku 4.5	1.00	5.00
Gemini 3.1 Pro Preview	2.00	12.00
Gemini 3.5 Flash	1.50	9.00
Gemini 3 Flash Preview	0.50	3.00
Gemini 3.1 Flash Lite	0.25	1.50
GPT-5.4	2.50	15.00
GPT-5.4-mini	0.75	4.50
GPT-5.4-nano	0.20	1.25
DeepSeek V4 Pro	1.74	3.48
DeepSeek V4 Flash	0.14	0.28
Gemma 4 31B-IT	0.13	0.38
Gemma 4 26B-A4B-IT	0.07	0.34
Qwen 3.6 27B	0.32	3.20
Qwen 3.6 35B-A3B	0.15	1.00

Table 11: Per-model list-price rate card (USD per million tokens) used to recompute costs in Table 12 and Figure 5. Rates are the published list-price standard tier, with no flex, batch, or prompt-caching discount applied; the same accounting basis is used for every model so cost comparison is consistent across providers and open-weight self-hosted models (open-weight prices are the published OpenRouter serving rate). Source: `runs/_or_pricing.json`.

Model	Naive RAG	Temporal RAG	FC (snap)	FC	Mean
Claude Opus 4.7	$ 0.5024	$ 0.4808	$ 0.3718	$ 0.3473	$ 0.4256
Claude Sonnet 4.6	$ 0.6254	$ 0.5334	$ 0.4545	$ 0.3222	$ 0.4839
Claude Haiku 4.5	$ 0.2898	$ 0.2487	$ 0.1456	$ 0.2029	$ 0.2217
Gemini 3.1 Pro Preview	$ 0.2218	$ 0.1654	$ 0.3200	$ 0.2971	$ 0.2511
Gemini 3.5 Flash	$ 0.6700	$ 0.5650	$ 0.6195	$ 0.5613	$ 0.6039
Gemini 3 Flash Preview	$ 0.0158	$ 0.0147	$ 0.0181	$ 0.0191	$ 0.0169
Gemini 3.1 Flash Lite	$ 0.0053	$ 0.0046	$ 0.0064	$ 0.0070	$ 0.0058
GPT-5.4	$ 0.1136	$ 0.1107	$ 0.1086	$ 0.0992	$ 0.1080
GPT-5.4-mini	$ 0.0390	$ 0.0413	$ 0.0261	$ 0.0283	$ 0.0337
GPT-5.4-nano	$ 0.0088	$ 0.0087	$ 0.0082	$ 0.0082	$ 0.0085
DeepSeek V4 Pro	$ 0.2341	$ 0.3664	$ 0.2870	$ 0.2332	$ 0.2802
DeepSeek V4 Flash	$ 0.0317	$ 0.0245	$ 0.0378	$ 0.0346	$ 0.0322
Gemma 4 31B-IT	$ 0.0028	$ 0.0024	$ 0.0035	$ 0.0039	$ 0.0031
Gemma 4 26B-A4B-IT	**$ 0.0014**	**$ 0.0013**	**$ 0.0021**	**$ 0.0025**	**$ 0.0018**
Qwen 3.6 27B	$ 0.0433	$ 0.0209	$ 0.0684	$ 0.0578	$ 0.0476
Qwen 3.6 35B-A3B	$ 0.0112	$ 0.0114	$ 0.0240	$ 0.0215	$ 0.0170

Table 12: Median list-price USD per item per model per retrieval setting on VerLexBench Mini, computed by re-applying the per-model rate card (Table 11) to the realized prompt and completion token counts of every run row. No prompt-caching, flex-tier, or batch-API discount is applied: every model is priced on the same standard-tier basis so the cost comparison is reproducible from token counts alone. Per-column minimum bolded. Self-hosted open-weight models are priced at their published OpenRouter serving rate.

Domain	Δ dense	Δ graph
Civil & Commercial	+0.367	+0.369
Tax & Revenue	+0.324	+0.275
Securities & Finance	+0.294	+0.312
Traffic & Transport	+0.286	+0.284
Labor	+0.212	+0.225
Land & Property	+0.208	+0.174
Consumer & Health	+0.190	+0.199
Criminal	+0.190	+0.196
Immigration	+0.184	+0.209
Environment	+0.184	+0.183
Const. & Election	+0.170	+0.157
Administrative	+0.144	+0.159
Intellectual Property	+0.128	+0.120
Family & Juvenile	+0.097	+0.073
Panel mean	+0.213	+0.210

Table 13: Per-domain temporal-misalignment cost (20 items per domain). Δ = panel-mean (Temporal $-$ Snapshot) TC-F2 lift, averaged over the sixteen-model panel within each domain. Dense = Snapshot RAG $\rightarrow$ Temporal RAG; graph = Snapshot FC $\rightarrow$ Temporal FC. Domains sorted by dense lift descending.