Do Small Models Use the Law You Give Them? Measuring Context Use on a Bilingual Bangladesh Legal Benchmark

Moniruzzaman Mahadi[1], Abrar Mohammed Tanzim Alam[1], Sayma Siddika Monalisa[1]
Mir Mohammad Asif Abdullah[1], Swakkhar Shatabda[2], Md Adnan Arefeen[1]

[1]Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh

Correspondence: momahadi9664@gmail.com

Abstract

When a legal question-answering system answers wrongly, accuracy alone cannot say why: the retriever may have missed the governing statute, the model may have ignored it, or the scorer may have misread the model's choice. We give a protocol that provides separate diagnostics for these three failures, and show that the scorer alone can manufacture a finding. On the 2022 and 2023 Bangladesh Bar Council examinations, a generate-then-parse pipeline credits fine-tuning with +50.0 points on one of six small instruction-tuned models where constrained option-letter scoring under all four rotations gives +3.0, and on one model reverses the sign; the gap tracks formatting, not legal knowledge. A supplied-law control that guarantees the governing provision among same-act distractors then isolates the model: five of six show rotation-invariant gains of 14.7 to 19.3 strict-consistency points, every interval excluding zero, the sixth only before rotation, and withholding that provision confirms the specific statute carries the effect. Fine-tuning changes the answer but, under single-provision training, not detectably its use of supplied law. We release the benchmark, control, corpus, adapters, per-item outputs, and analysis code. The dataset is publicly available at https://huggingface.co/datasets/momahadi/bangladesh-legal-qa-dataset.

1 Introduction

A retrieval system can put the right statute in front of a language model and the model can still answer the question wrongly. The retriever may have returned the wrong provision, the model may have failed to use the right one, or the scorer may have misread what the model selected, and end-to-end accuracy cannot tell them apart. Figure 1 shows how we separate the three.

One response is to fine-tune the model on examples containing the governing provision. We test this on six small instruction-tuned models from three families, in Bangla and machine-translated English, on 199 items from the 2022 and 2023 Bangladesh Bar Council examinations.

Measuring this is harder than it looks. Small models often fail to produce the answer format a script expects, so a strict parser scores formatting rather than law, and fine-tuning fixes formatting quickly. In our data a strict parser credits fine-tuning with +50.0 points on one model where matched constrained scoring gives +3.0, and on another it reverses the sign. We read the answer as an argmax over the four option-letter logits under all four option rotations (Pezeshkpour and Hruschka, 2024; Zheng et al., 2024) and count an item correct only when all four agree.

Our results are conditional. Supplying the governing provision raises strict consistency by 14.7 to 19.3 points for five of six models. Removing only that provision while keeping same-act text costs 8.0 to 15.3 points, every interval excluding zero. The exception is the smallest, which does not improve when handed the provision it needs. Fine-tuning helps weak models but not strong ones and, under single-provision training, does not detectably change use of supplied law. The two evaluation sets disagree, so we do not claim fine-tuning teaches context use. Constrained scoring measures which answer the model prefers, not how it reasons; the supplied-law control is

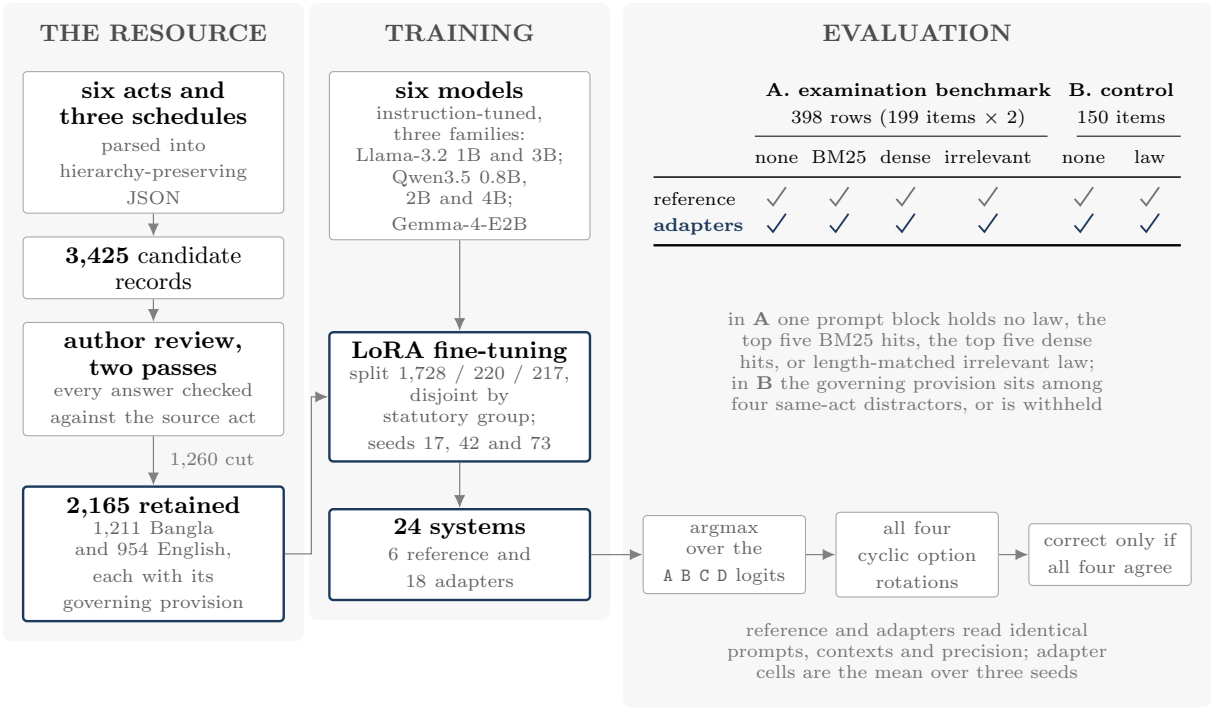

Figure 1: Study design. The retained records train the adapters, and every resulting system is read under the same items, contexts and scorer. A tick marks a condition that arm was run under.

what makes that preference a test of law use. What generalizes from six models is the protocol and the artifacts it exposes, not a scaling law.

This paper contributes a matched application of constrained option-letter scoring with full cyclic rotation (Liu et al., 2024; Robinson et al., 2023), whose nulls and reversals show that the scorer decides the finding; a supplied-law control that separates model failure from retriever failure; and a bilingual Bangladeshi legal evaluation set with the measurement protocol that connects them. We release the inputs, adapters, per-item outputs, manifests, and analysis. The task is answer selection, a proxy for legal knowledge and not a substitute for legal practice.

2 Related Work

Bangladesh and Bangla legal QA. MINA evaluates a bilingual, tool-using legal assistant on the same 2022 and 2023 Bangladesh Bar Council examinations used here (Wasi et al., 2026). Its strongest MCQ results come from a proprietary model with two-stage retrieval or tools and are five-run averages under one option order, the single-order scoring that Section 6.1 shows can move a legal-QA result by tens of points, so they are not comparable to our strict four-rotation score. The retrieval pools also differ: MINA indexes 595 acts and 18,023 sections from the national portal, ours the six acts and three schedules of the Bar Council syllabus. Its indices and data are promised on acceptance and were unavailable during this work, so we cannot score our systems under its retrieval. A separate study collects 250 Bangla questions from a public legal-help forum and evaluates four models, with its detailed methods reporting professional-lawyer validation on a 50-question subset (Aftahee et al., 2025). It reports well-structured answers alongside fabricated citations and unsafe advice. These studies establish the local task and its risks, but do not isolate whether a wrong answer follows from missing law, failure to use supplied law, or the evaluation instrument.

English legal benchmarks are broader: LexGLUE standardizes seven NLU datasets including five-choice CaseHOLD (Chalkidis et al., 2022), and LegalBench collects 162 professionally written tasks across six kinds of legal reasoning (Guha et al., 2023). We test

statute-based answer selection, not legal reasoning.

Retrieval and supplied context. LegalBench-RAG evaluates retrieval separately and favors minimal, precisely relevant legal spans over broad chunks (Pipitone and Alami, 2024). NitiBench reports retrieval and end-to-end results for Thai law, including cases where added relevant sections rank too low to help the generator (Akarajaradwong et al., 2025). Context itself can alter what a model recovers: in post-cutoff United States opinions, recall follows a J-shaped curve as local coherence is disrupted (Xia et al., 2025). These motivate component-level evaluation. Our supplied-law control holds retrieval success fixed instead: the governing provision is present by construction, so a remaining error belongs to the model side.

Evaluation validity. Even closed-form MCQ scoring is not automatically stable: reordering options can move accuracy by 13 to 85 points (Pezeshkpour and Hruschka, 2024), and prior probability on option-ID tokens introduces selection bias (Zheng et al., 2024). More generally, plausible prompt formats can reverse model comparisons (Sclar et al., 2024), small changes to option order can move leaderboard rankings by up to eight places (Alzahrani et al., 2024), and exact task implementations are part of a reproducible result (Biderman et al., 2024). GreekBarBench, which validates LLM judges against expert annotations on open-ended bar-exam questions, illustrates the broader need for scorer-level scrutiny in legal QA (Chlapanis et al., 2025). Held and Habernal (2025) likewise find that fine-tuning lifts output format validity from 0.620 to 0.992 while a domain-pretrained encoder still wins the task itself, the same format-versus-knowledge split Section 6.1 measures. The established remedies are to require correctness under every cyclic option shift (CircularEval, introduced with MMBench (Liu et al., 2024)) and to select from option-ID logits rather than free generation (Robinson et al., 2023), avoiding the surface-form competition that Holtzman et al. (2021) show misranks options independently of what the model prefers. We adopt that combination unchanged. The new question is what this established instrument does to a legal adaptation result when prompts, contexts, models, and runtime are held fixed.

3 Data and Evaluation Sets

Statutory corpus. The corpus covers six acts central to the Bangladesh Bar Council syllabus: the Penal Code 1860, Code of Criminal Procedure 1898, Evidence Act 1872, Code of Civil Procedure 1908, Specific Relief Act 1877, and Limitation Act 1908. It also includes CPC Schedule I, CrPC Schedule II, and Limitation Act Schedule I. English text comes from the official Bangladesh law portal (`http://bdlaws.minlaw.gov.bd/`). The conversion preserves the hierarchy from act to clause and separates statutory illustrations from clauses that use the same parenthesized labels (Appendix A).

The three schedule files are available only in English in our corpus. When a Bangla training question is governed by a schedule, the question and options remain Bangla but the supplied schedule text remains English. We use the same convention at evaluation. Thus the language label describes the question, not necessarily every legal passage in its prompt.

Fine-tuning records. Question generation produced a preserved pool of 3,425 records. Two review passes retained 2,165: 1,211 Bangla and 954 English, with 795 single-hop, 712 advanced-selection, and 658 bar-exam-style records. The authors checked each retained answer against its cited provision and removed incomplete, repetitive, or cross-act items. The complete fine-tuning set did not receive legal-expert review. The 658 multiple-choice records use a one-line answer target that names the chosen option inline in prose, never the `FINAL_ANSWER: (` stem the constrained scorer appends at evaluation. The remaining records use their prose answers.

Each retained record stores the governing provision and, for some generation tasks, a candidate pool under `Possible Sections`. The training conversion does not use that candidate pool. It places only the governing provision in `LEGAL_CONTEXT`. In the 1,728-row training split, this context has a median length of 209 characters and only four rows exceed 1,500

characters. This fact matters for the intervention: training supplies clean law, while retrieval evaluation supplies five candidates with a median rendered length of 2,002 characters.

3.1 Bar Council Benchmark

The benchmark contains the 2022 and 2023 Bangladesh Bar Council MCQ examinations. The original Bangla questions and official answer keys are paired with English translations of the same items. Item 2022-005 was cancelled by the examining body and is excluded in both languages. The final benchmark therefore has 199 underlying questions and 398 language rows. Repeated languages, contexts, rotations, and seeds do not increase the independent item count.

The English side was machine translated and was not checked by a legal expert. We audited all 199 pairs with two model judges from a different vendor than the translator, using gross option substitutions rather than subtle terminology errors; the weaker judge caught only 3 of 16. All 29 flags from the stronger judge lie in the 2022 paper, yet 2022 has the smaller context-specific effect for five of six models. Excluding flagged items moves the effects by 0.8 to 2.1 points and reverses no sign (Appendix G). This bounds one class of translation error; it is not expert review.

The 2022 and 2023 papers were withheld from question generation. A same-language comparison between each benchmark stem and every training question finds no exact match: the maximum SequenceMatcher ratio is 0.717 and the maximum character 5-gram Jaccard 0.463. That 0.717 pair is a bar-exam-style Limitation Act schedule question and benchmark item 2022-086, both testing the limitation period for specific performance, so there is no verbatim reuse but there is substantive overlap. These surface checks measure neither provision-level overlap nor whether a public model saw the examinations during pretraining.

3.2 Supplied-Law Control

The control contains 150 new MCQs, balanced across Bangla and English and across the six acts. Each item supplies five provisions from the same act, one of which governs the answer. The rendered context has median length 1,643 characters and maximum length 1,948, within the benchmark's 2,000-character retrieval budget. No-law heuristics (single-order plain accuracy, not strict consistency) give 26% for the longest option, 25% for the longest unhedged option, and 37% for the most-quoted option, against 25% chance.

GPT-5.6-Luna drafted both the English half (`handbuild-en-v1`) and the Bangla half (`handbuild-bn-v1`); it does not belong to an evaluated model family, and no evaluated-model score was viewed before the control was frozen. Two authors then checked every question, all four options, and the answer key against the governing provision; 20 items were discarded. After all experiments were complete, a law graduate and a law masters student independently confirmed every answer against its cited provision without seeing model outputs; no items were changed (Appendix K). When the governing provision is guaranteed present, retrieval failure cannot explain a wrong answer.

The released control preserves the context strings used in inference. As with training (Section 3), schedule text remains English even in Bangla items; 14 of 75 Bangla items contain at least one majority-English passage. Reference and adapter systems always receive the same string.

4 Method

We ask three questions. Does a generation-and-parser scorer change the measured effect of fine-tuning? Can each model use law that is guaranteed present? And does context-injected fine-tuning change how much that law helps? The first two questions establish the measurement needed for the third.

Intervention and matched systems. The intervention is a LoRA adapter (Hu et al., 2022) trained on examples whose prompt contains the governing provision. For each model, the *reference* is the released instruction-tuned weights without an adapter. Three adapters are trained independently with seeds 17, 42, and 73. Within a comparison, reference and adapter use the same model revision, chat template, prompt, context string, precision, runtime, and scorer. The adapter is attached at inference and is never merged.

The six models are Llama-3.2-1B-Instruct and Llama-3.2-3B-Instruct (Meta AI, 2024); Qwen3.5 at 0.8B, 2B, and 4B (Qwen Team, 2026); and Gemma-4-E2B-it (Gemma Team, 2026). Qwen3.5 uses FP16 LoRA. Llama and Gemma use NF4 QLoRA (Dettmers et al., 2023) to fit a 16 GB accelerator. Precision is matched within every reference-adapter pair, so it cannot explain a within-pair difference. It can still contribute to cross-family differences.

Training and evaluation contexts. Training supplies one governing provision in isolation. It does not present the five-candidate pools stored during data generation, in part because five candidates would roughly triple the median prompt length and exceed the 1,536-token training sequence length we use to fit a single T4. Evaluation is therefore a harder discrimination setting: BM25 and dense retrieval each supply the top five sections, capped at 2,000 characters, while the control supplies one governing provision among four same-act distractors. Keeping training context clean leaves distractor handling to evaluation instead of training it into the adapter.

The examination benchmark has three primary context conditions. **none** supplies no statute. **BM25** supplies the top five lexical hits (Robertson and Zaragoza, 2009). **dense** supplies the top five hits from a frozen BGE-M3 encoder (Chen et al., 2024). Retrieval runs once within each language pool, and the resulting strings are reused for every system. The benchmark carries official answer keys but no expert labels for which statutory provisions govern each item, so these are retrieval conditions rather than verified relevant-law conditions.

The frozen benchmark protocol also included shuffled irrelevant context. After the run, we found that it was much shorter than retrieved context, so we replaced it with a length-matched control: five same-language sections excluded from that item's BM25 and dense hits, with median length 2,000 characters against 2,002 for retrieval. We use it only for the relevance-versus-length diagnostic. The original retrieval-minus-none contrast is unchanged.

A separate 150-item supplied-law control (Section 3.2) tests whether models use law that is guaranteed present. Its withheld-provision pass (Section 6.2) then replaces the governing provision with same-act text to confirm the provision carries the effect. Questions, options, keys, candidate pools, prompts, and analysis were fixed before any output was examined. Each run carries a 52-row replication canary that re-runs the **none** condition on a stratified benchmark sample and gates on identical strict outcomes, verifying that the session reproduces the frozen main run before any new result is read.

Constrained scoring and option rotation. The primary scorer is `constrained-letter-logit-v1`, an instance of the option-ID scoring recommended by Robinson et al. (2023). The frozen prompt is rendered with the model's own chat template, the literal stem `FINAL_ANSWER: (` is appended, and one forward pass produces logits for the single-token labels `A`, `B`, `C`, and `D`. The argmax is the prediction. This path is deterministic, cannot produce a parse failure, and does identical work for reference and adapter. Each item is evaluated under all four cyclic option rotations. The gold label moves with its option, and an item counts as correct only if all four predictions are correct. This is the CircularEval criterion of Liu et al. (2024), applied here to four options. A uniformly random system scores $(1/4)^4 = 0.39\%$, and a fixed-letter strategy scores zero. We refer to the resulting measure as rotation-invariant strict consistency and use it throughout unless stated otherwise.

Parser diagnostic. For the dense condition at rotation zero, we also decode a free-form answer and apply a strict parser that requires one line matching `FINAL_ANSWER: ([ABCD])`. The diagnostic pools the 199 Bangla and 199 English rows. It reads the same reference and seed-42 systems in two ways: generated text through the parser, and option-letter logits through the primary scorer. We use one option order because this diagnostic reproduces the common single-order generation-and-parser setting; it does not enter the primary outcome. Free-form answers were retained only for rotation zero, so extending the parser comparison across rotations would re-

Model	reference (%)			FT mean (%)			FT − ref (95% CI)		
	none	law	w/o	none	law	w/o	none	law	w/o
Llama-3.2-1B	28.7	27.3	24.0	**44.9**	41.8	35.3	+16.2 [+10.4, +22.4]	+14.4 [+7.6, +21.3]	+10.7 [+4.2, +17.3]
Qwen3.5-0.8B	38.7	57.3	49.3	52.0	**71.8**	57.8	+13.3 [+7.8, +19.3]	+14.4 [+8.0, +21.1]	+8.4 [+2.0, +15.1]
Llama-3.2-3B	50.7	68.0	58.0	53.8	**70.2**	55.3	+3.1 [−2.2, +8.4]	+2.2 [−4.7, +8.9]	−2.7 [−9.1, +3.6]
Qwen3.5-2B	57.3	76.7	64.0	60.0	**78.4**	63.6	+2.7 [−2.0, +7.3]	+1.8 [−3.8, +7.3]	−0.4 [−6.9, +6.2]
Gemma-4-E2B	70.0	**84.7**	69.3	58.9	78.7	64.4	−11.1 [−17.6, −4.9]	−6.0 [−11.3, −0.7]	−4.9 [−9.8, −0.2]
Qwen3.5-4B	76.7	**94.0**	79.3	69.1	91.1	77.1	−7.6 [−13.6, −1.8]	−2.9 [−6.9, +1.1]	−2.2 [−6.9, +2.4]

Table 1: Strict consistency on the 150-item control, in percent. **law**: governing provision present among five same-act candidates. **w/o**: governing provision replaced with same-act text (Section 6.2). **FT** columns are three-seed means. **FT − ref** is the fine-tuning effect (adapter minus reference) under each condition, with 95% item-clustered bootstrap intervals over the 150 items. Bold: highest raw score per row.

quire new inference.

Protocol integrity. The model matrix, prompts, training recipe, primary scorer, and statistical plan were frozen in a written contract before the examination benchmark was opened. Run manifests record the contract and prompt SHA-256 digests. Every follow-up run also passes a replication canary against frozen benchmark rows. The missing oracle condition and the length-matched replacement are disclosed as deviations from the original protocol. The supplied-law control is a separate experiment with its own pre-specified design.

5 Experimental Setup

Split and training. The 2,165 records are split by normalized (act, section) group so that a statutory group appears in only one partition: 1,728 training, 220 validation, and 217 internal test records over 749, 105, and 98 groups. The released split includes its group-assignment digest. The Bar Council benchmark is not used for training, prompt design, early stopping, or hyperparameter selection.

All models use LoRA rank 16, alpha 16, all linear language-model projections, effective batch size 16, at most two epochs, and early stopping on validation loss. Learning rate is selected from the frozen grid $\{5e{-}5, 1e{-}4, 2e{-}4\}$ by a one-epoch pilot on a fixed 512-record subset. Every model selects $2e{-}4$, the upper boundary; this is also the default recommended by the Unsloth framework (Han et al., 2023). We did not widen the grid after observing that result. Each seed is trained on one 16 GB T4.

Aggregation and uncertainty. The statistical unit for the examination benchmark is the original question. The two language versions, four contexts, four rotations, and three adapter seeds are repeated measurements, not independent observations. Tables report the reference score and the mean across adapter seeds. Confidence intervals use 10,000 item-clustered bootstrap draws. The paired tests are two-sided exact McNemar tests on item outcomes. Holm correction is applied within the frozen family of 12 tests per model: two languages, two retrievers, and three seeds.

Context specificity is a difference in differences. For context C,

$$\big(\mathrm{FT}_C - \mathrm{FT}_{\mathrm{none}}\big) - \big(\mathrm{Ref}_C - \mathrm{Ref}_{\mathrm{none}}\big).$$

For the examination summary, C is the mean of BM25 and dense within each seed. The supplied-law control uses supplied law for C. A positive value says fine-tuning helps more when that context is present. It does not by itself say that the adapter is better than the reference under context.

The bootstrap intervals summarize estimation uncertainty and are not adjusted for the many contrasts shown in the tables. We treat counts of intervals that exclude zero as descriptive. Confirmatory multiplicity control applies only to the pre-specified McNemar family described above.

6 Results

Table 1 summarizes the supplied-law control and Table 2 the examination results. We read the scorer first, because it decides what both tables mean.

	none		BM25		dense		irrelevant		context specificity		
Model	Ref	FT	Ref	FT	Ref	FT	Ref	FT	R−N	R−I	Holm
Bangla											
Llama-3.2-1B	0.5	**5.5**	0.5	**8.5**	2.0	**13.1**	0.5	**6.0**	+4.5 [+0.9, +8.1]	+4.0 [+0.0, +8.2]	6/6
Qwen3.5-0.8B	3.5	**4.0**	11.6	**17.3**	19.6	**24.1**	6.0	**7.4**	+4.6 [+1.2, +8.0]	+3.8 [−0.4, +7.9]	1/6
Llama-3.2-3B	**6.5**	5.7	12.1	**17.1**	14.1	**23.6**	4.0	**8.4**	+8.1 [+3.7, +12.6]	+2.9 [−1.8, +7.7]	3/6
Qwen3.5-2B	2.5	**6.4**	19.1	**20.8**	23.1	**26.8**	6.5	**7.2**	−1.2 [−5.5, +3.0]	+2.0 [−2.2, +6.1]	0/6
Gemma-4-E2B	**8.5**	7.5	**27.1**	26.3	**37.7**	32.8	**11.1**	6.4	−1.8 [−6.2, +2.6]	+1.8 [−2.1, +5.9]	0/6
Qwen3.5-4B	9.5	**11.7**	24.1	**27.6**	28.6	**36.5**	**11.1**	10.9	+3.5 [−0.3, +7.2]	+5.9 [+1.8, +10.2]	2/6
English											
Llama-3.2-1B	8.5	**15.4**	11.6	**25.5**	15.6	**34.8**	4.0	**10.2**	+9.7 [+4.3, +15.2]	+10.4 [+5.1, +15.5]	6/6
Qwen3.5-0.8B	11.6	**13.2**	24.6	**33.2**	31.2	**41.4**	7.0	**11.2**	+7.7 [+3.1, +12.3]	+5.2 [+0.1, +10.2]	6/6
Llama-3.2-3B	**20.1**	19.1	34.7	**37.5**	**43.7**	42.9	17.6	**18.8**	+2.0 [−3.4, +7.3]	−0.2 [−5.3, +4.9]	0/6
Qwen3.5-2B	14.6	**20.8**	36.7	**38.7**	38.7	**45.4**	13.6	**15.9**	−1.8 [−6.9, +3.3]	+2.0 [−3.2, +7.3]	0/6
Gemma-4-E2B	21.1	21.1	**37.7**	35.8	**49.7**	45.2	**19.6**	15.6	−3.2 [−7.5, +1.3]	+0.8 [−3.6, +5.4]	0/6
Qwen3.5-4B	28.1	**29.6**	**44.2**	41.2	50.8	**53.1**	**25.1**	24.0	−1.8 [−7.0, +3.1]	+0.8 [−3.3, +5.0]	0/6

Table 2: Strict consistency on the 199-item examination (%). **FT**: three-seed mean. **R−N**, **R−I**: retrieval minus no-context and minus irrelevant. **Holm**: surviving McNemar tests of six. Bold marks the higher value, not significance; +0.0 does not exclude zero. Sparse cells: Table 9.

6.1 The Scorer Changes the Measured Effect

We read the same 398 dense, rotation-zero outputs three ways. For Qwen3.5-2B, an exact-line parser gives the reference 4.0% accuracy, a lenient parser gives 33.2%, and constrained letter logits give 51.0%. Its seed-42 adapter scores 54.0% under all three readings. The apparent fine-tuning gain is therefore +50.0, +20.9, or +3.0 points. The scorer changes what we conclude about fine-tuning, not what the model selected. The scorer can also change which system wins. For Gemma-4-E2B, the parser gives the adapter a +10.1-point advantage, while constrained scoring gives −2.5. This sign reversal holds across all three seeds. We report it as a warning about the scorer, not as a finding about Gemma-4-E2B, whose adapter also shows signs of over-adaptation. By contrast, Qwen3.5-4B is 98.7% format-compliant and the two effects differ by less than one point. Across all six, the scorer gap tracks format compliance: 47.0 points at 5.8% to 0.8 at 98.7% (Table 10), so the artifact is formatting, not knowledge. Option order adds a second distortion, spanning 13.6 to 57.3 points across gold positions on Bangla dense rows, so all remaining results use constrained scoring under all four rotations.

6.2 Five Models Benefit from Supplied Law

The **none** and **law** columns of Table 1 compare each reference on the same 150 items with and without the governing provision. Five models gain 14.7 to 19.3 points, each confidence interval excluding zero, and Qwen3.5-4B reaches 94.0% overall and 100% on the English half. These models can use the law when it is guaranteed present.

Copying the most text-matched option does not explain this. On the 92 items where another option has more 5-gram coverage of the supplied law, the same five models gain 12.0 to 22.8 points and Llama-3.2-1B is again the exception at −2.2.

Llama-3.2-1B is the exception. Its score changes from 28.7% without law to 27.3% with law, a −1.3-point difference with interval $[-7.3, +4.7]$, which we read as no detected gain rather than harm. Retrieval cannot explain it, since the provision is present by construction. At rotation zero it gains 6.7 points from supplied law and loses that only once all four rotations must agree (Appendix I): position instability, not an inability to use the provision. It is also the oldest model (released 2024 against 2026), so generation is confounded with size; Llama-3.2-3B bounds that reading with the same family and release and a +17.3-point law effect.

The withheld-provision pass replaces only the governing provision with another same-

act passage of matched length, keeping the other four distractors. The replicated gold and none arms reproduce on 1,798 of 1,800 strict outcomes; the two Bangla items that differ are argmax ties in one session, which moves the within-session law effect by 0.7 points on the two Llama rows. Removing the provision costs the five responsive models 8.0 to 15.3 points, every interval excluding zero. Same-act text without it is worth −0.7 to +10.7, and only two of those intervals exclude zero. The provision carries the law effect, although for Qwen3.5-0.8B same-act text contributes more of it (43%).

6.3 Change in Strict Consistency Varies with Starting Accuracy

The reference **none** column gives a common starting point. The two models below 40% gain 13.3 and 16.2 points, the two in the middle gain 2.7 and 3.1, and the two above 70% lose 7.6 and 11.1. With only six models, this is a descriptive pattern, not a law of model size or family.

Supplying the governing provision does not create a consistent extra fine-tuning benefit. For the four references below 60%, the effects with and without law differ by at most 1.8 points, and the two strongest references decline in both conditions. The control differences in differences run from −1.8 to +5.1 with wide intervals (Table 8), bounding this effect rather than resolving it. Fine-tuning mainly changes answer selection rather than use of supplied law.

6.4 Absolute Scores Are Lower in Bangla

Absolute scores are lower in Bangla (Table 2), but the control's law effect is not: the five responsive models gain 13.3 to 21.3 points in Bangla against 12.0 to 21.3 in English, with Qwen3.5-4B at 88.0% against 100%. Llama-3.2-1B again differs, at −4.0 against +1.3. The absolute gap is confounded with translation quality; the difference in differences is not (Appendix G).

6.5 The Two Evaluation Sets Differ in Resolution

The examination benchmark gives a different answer for three models in Table 2. Its retrieval-minus-none difference in differences is +7.1 for Llama-3.2-1B, +6.2 for Qwen3.5-0.8B, and +5.1 for Llama-3.2-3B. Five of the twelve descriptive R−N intervals exclude zero, all of them positive, as do three of the R−I intervals.

Nine of the 96 cells sit at or below 5% (Table 9), so the floor limits what those contrasts can say. The floor alone does not explain the gap: the likeliest cause is training-evaluation mismatch, since adapters saw one clean provision but must select among five candidates at test.

7 Discussion

The supplied-law control changes how retrieval results should be read. Five models improve when the governing provision is present; the withheld-provision pass confirms the provision carries the majority. Llama-3.2-1B is the exception, so its weak end-to-end result cannot be assigned to retrieval alone. The examination set and the control then disagree because context specificity depends on the test distribution, not only the adapter: the adapters never learned to select governing law from five candidates.

The fine-tuning pattern does not identify a mechanism: null on law use, mixed on accuracy, and bounded by ceiling effects. Family, size, generation, and precision remain alternatives because the six are not a controlled series.

Measurement is the guardrail. A parser can reward format learning even when answers barely change, as the Gemma reversal shows; constrained logits and rotations make the comparison defensible but do not measure explanation quality, which still needs expert or validated-judge rubrics (Chlapanis et al., 2025).

8 Conclusion

Five of six small instruction-tuned models use supplied Bangladeshi statute in Bangla and machine-translated English: on a constructed control, the governing provision raises strict consistency by 14.7 to 19.3 points, and removing only that provision costs 8.0 to 15.3, every descriptive interval excluding zero. A parser would have reported +50.0 where logit scoring gives +3.0, and reversed

a sign. Single-provision fine-tuning does not detectably change use of supplied law. Llama-3.2-1B uses the law too, but only while the options stay still. Training on five-candidate prompts remains untested.

Limitations

Scope and coverage. The benchmark has 199 underlying questions from two examination years in one jurisdiction. Bangla and English, contexts, rotations, and seeds are repeated measurements of those questions. Multiple-choice answer selection is a proxy for legal knowledge, not legal reasoning or practice. Strict consistency also penalizes option-order instability, so a low score can reflect brittleness as well as a wrong preferred answer.

Training supplies one governing provision, while retrieval and control evaluation supply five candidates. This is a deliberate transfer test: whether learning from isolated governing provisions carries over to selecting among five candidates, not the stronger intervention of training directly on distractor-rich contexts. Repeating the six-model, three-seed grid on that stronger setup requires longer sequences and more training than the single-T4 budget used here, so we scope it as defined future work rather than a limit on what fine-tuning can do.

All models select the upper boundary of the learning-rate grid. A wider search could change effect size or direction. Qwen3.5 uses FP16 LoRA, while Llama and Gemma use NF4 QLoRA. Precision is matched within each comparison but not across families. The models are not a controlled scaling series, so the relation with starting accuracy is descriptive. Public models may also have seen the public examinations during pretraining. One training item tests the same Limitation Act rule as benchmark item 2022-086 with a SequenceMatcher ratio of 0.717, despite no exact match. Repeated development against the same 199 items creates further co-adaptation risk.

Validity of the resource. The Bangla benchmark questions and keys come from the Bangladesh Bar Council. The English versions are machine translations and received no legal-expert review. Two model judges provide a measured sensitivity analysis, not expert validation. The 2,165 fine-tuning records were checked against their cited provisions by the authors but were not exhaustively reviewed by a lawyer. The 150 control items were drafted by GPT-5.6-Luna, which is outside the evaluated families. Two authors checked every question, option set, and key against the governing provision, and 20 items were removed. They are not expert-authored.

The statutory schedules in the released corpus are English-only. Bangla questions governed by those schedules therefore retain English legal context, as they do in the training set. In the control, 14 of 75 Bangla items contain at least one majority-English candidate passage. This limits language-specific interpretation of the control. It does not break the matched reference-adapter comparison because both systems receive the identical recorded context.

The benchmark has no expert labels for which provisions govern each item. We cannot report oracle retrieval or Recall@k, and a language-specific retrieval difference could appear as a language-specific adaptation effect. Hashing and reusing contexts ensures that reference and adapter receive identical evidence. It does not establish that the evidence is legally sufficient. The supplied-law control closes this gap only for its 150 constructed items.

Validity of the measurement. The inflation result rests on cleanly trained models. The sign reversal rests on Gemma-4-E2B alone, which carries three warning signs together: the only large gap between final training and validation loss (0.132/0.875), the only Bangla position spread that grows after fine-tuning (+4.8 points), and opposite-signed parser and logit effects. That is consistent with over-adaptation and changed answer-surface behavior but proves neither, and no lower-rate adapter was scored. Whether the reversal is a property of the scorer or of an over-adapted adapter is not settled here, so the two findings should be weighted differently.

The control's law effect contrasts no context against on-topic statutory text containing the answer. A follow-up withheld-provision pass (Section 6.2) decomposes it, but that decomposition covers reference systems only and was not part of the frozen contract.

Table 10 shows the scorer gap shrinking as reference format compliance rises, from 47.0 points at 5.8% compliance to 0.8 at 98.7%. A one-shot format demonstration in the system message might recover much of that gap without any fine-tuning. We did not test it, so we cannot say how much of the measured inflation is a property of the scorer and how much is a property of unprimed zero-shot references.

The length-matched irrelevant context replaced a shorter frozen control after the length mismatch was observed. We report it as a sensitivity condition and keep retrieval minus no context as the frozen benchmark contrast. The supplied-law control is a separate experiment whose data, prompts, and analysis were fixed before its outputs were examined. It was not part of the original benchmark contract.

Ethics Statement

Legal question answering is a sensitive task. We treat every system as a research artifact and make no claim that it can provide legal advice. The strongest configurations still fail many items under strict scoring, and the task omits drafting, citation verification, client facts, procedural judgment, and professional responsibility. A bar-exam score must not be read as evidence of competence to advise a person.

The released material contains public statutes and public examination papers, with no client records or personal data. Our generated questions, answers, annotations, translations, metadata, and arrangement are released under CC BY 4.0 only to the extent that we hold rights in them. We do not claim ownership over official statutory text or examination questions. The anonymous supplement removes author names, personal links, and identifying metadata for double-blind review.

The main benefit claimed here is safer evaluation practice. Reporting an adapter gain from a parser that rewards format compliance can misstate legal-QA progress. Separating answer selection from formatting, and model failure from retriever failure, gives reviewers and future users a more accurate boundary on what the experiment supports.

References

Sabik Aftahee, A. F. M. Farhad, Arpita Mallik, Ratnajit Dhar, Jawadul Karim, Nahiyan Bin Noor, and Ishmam Ahmed Solaiman. 2025. Assessing the reliability of large language models in the Bengali legal context: A comparative evaluation using LLM-as-judge and legal experts. *Preprint*, arXiv:2511.05627.

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. Association for Computational Linguistics.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, et al. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 13787–13805.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, et al. 2024. Lessons from the trenches on reproducible evaluation of language models. *Preprint*, arXiv:2405.14782.

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*, pages 4310–4330.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335.

Odysseas S. Chlapanis, Dimitrios Galanis, Nikolaos Aletras, and Ion Androutsopoulos. 2025. GreekBarBench: A challenging benchmark for free-text legal reasoning and citations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25099–25119. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems 36*.

Gemma Team. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, et al. 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36.

Daniel Han, Michael Han, and Unsloth Team. 2023. Unsloth. Software repository, version 2026.7.6 as recorded in the run manifests. Accessed 2026-07-16.

Lena Held and Ivan Habernal. 2025. Contemporary LLMs struggle with extracting formal legal arguments. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 292–303. Association for Computational Linguistics.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7038–7051. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations (ICLR)*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. MMBench: Is your multimodal model an all-around player? In *Proceedings of the European Conference on Computer Vision (ECCV)*. Oral presentation. Introduces CircularEval.

Meta AI. 2024. Llama 3.2 model card. `https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md`. Official model card for the Llama 3.2 1B and 3B releases.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.

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

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. Qwen Blog. Official model announcement. Accessed 2026-07-17.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Azmine Toushik Wasi, Wahid Faisal, Mst Rafia Islam, and Md Rizwan Parvez. 2026. Mina: A multilingual LLM-powered legal assistant agent for empowering access to justice in Bangladesh. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 25980–26028. Association for Computational Linguistics.

Eric Xia, Karthik Srikumar, Keshav Karthik, Advaith Renjith, and Ashwinee Panda. 2025. Beyond the haystack: Sensitivity to context in legal reference recall. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 48–53. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations (ICLR)*.

A Corpus Construction

Sources and scope. English statutory text comes from the official Bangladesh law portal maintained by the Ministry of Law, Justice and Parliamentary Affairs (http://bdlaws.minlaw.gov.bd/); Bangla text comes from published act collections. The same Bangla renderings appear across several published editions, so we treat them as the standard published text rather than any one publisher's. Scope follows the Bar Council syllabus rather than the statute book: the Penal Code 1860 (Act XLV of 1860), Code of Criminal Procedure 1898 (Act V of 1898), Evidence Act 1872 (Act I of 1872), Code of Civil Procedure 1908 (Act V of 1908), Specific Relief Act 1877 (Act I of 1877), and Limitation Act 1908 (Act IX of 1908), together with CPC Schedule I, CrPC Schedule II, and Limitation Act Schedule I.

The structure problem. Bangladeshi statute nests act, part, chapter, section, subsection, clause, and subclause. Sections are numbered with Arabic numerals and a full stop (`1.`), subsections with parenthesised Arabic numerals (`(1)`), clauses with parenthesised lowercase letters (`(a)`), and subclauses with parenthesised Roman numerals (`(i)`). Illustrations and explanations may attach at any level from section to subclause, and illustrations reuse the parenthesised lowercase form that clauses already use, so `(a)` is ambiguous on its face. Illustration sequences are, however, always introduced by an explicit label, which makes the two separable by a regular-expression pass before any model sees the text. Running that pass first is what allows the converted JSON to keep illustrations attached to the provision they illustrate instead of flattening them into clause lists.

Conversion. Each act was converted to JSON with a schema-enforcing prompt, one per act because formatting conventions differ between them and no single prompt held across the set. The schema fixes a root object carrying the act name, enactment date, preamble, and either parts or chapters; each section carries its number, marginal note, text, and optional subsections, clauses, and illustrations; each subsection and clause recurses in the same shape. The prompt also specifies act-specific cleaning, for example ignoring page headers, repeated running titles, and line numbers, merging lines broken mid-sentence, recovering marginal notes that appear before, beside, or interleaved with section numbers, and recording omitted or repealed sections rather than dropping them. The prompt was revised until the output conformed, and the files were then inspected by hand.

A converted section keeps its full hierarchy. Section 81 of the Penal Code, for instance, retains its chapter (GENERAL EXCEPTIONS), its marginal note, its section text, its explanation, and both illustrations as separate labelled objects rather than as one text blob, which is what makes it possible to supply a provision to a model at inference time without also supplying the illustration numbering as noise.

B QA Generation and Review

Categories. Three categories were generated for each language, with a dedicated prompt for each of the six combinations. **Single-hop extraction** asks for an answer contained in one supplied provision, and its prompt supplies that provision alone. **Advanced selection** supplies a list of exactly five candidate provisions, the governing one and four distractors, all copied verbatim from the corpus, and requires elimination. **Bar-exam** items are written in the style of examination questions: the prompt is supplied with a reference collection of bar-examination-style questions, drawn from examinations other than the 2022 and 2023 papers, and produces items in that form paired with the governing provision and four distractors located in the corpus. It also carries a `Cancelled` flag so that voided reference questions are marked rather than silently dropped. The answer keys in this category are therefore assigned during generation and checked in review, not inherited from an examining body; that distinction matters because the benchmark's keys are inherited and this category's are not.

Where the governing provision sits. Both selection prompts require the correct candidate to move. The instruction is that its position, first through fifth, "must vary randomly across entries, with approximately equal frequency," and must never default to first. The instruction was followed only in part. Across the 712 advanced-selection records the governing provision falls at position 1 in 13.8% of entries, 2 in 50.4%, 3 in 23.7%, 4 in 7.0% and 5 in 5.1%, so it never defaults to first but is far from uniform. There is no fixed answer slot to learn, and there is a positional tendency we did not intend. This matters because the reference models carry strong positional priors (Section 6), and it is one more reason the primary metric scores every benchmark item under all four option orders rather than one.

Separation for the bar-exam category. This is the category whose prompt sees examination-style reference material, so it carries whatever leakage risk exists. Three things bound it. The reference collection is drawn from examinations other than the two we evaluate on, and the 2022 and 2023 papers were never supplied to any generation prompt. The reference items acted as style exemplars during generation and were not carried into the released records. Every retained record then passed two author review passes against the source law. Comparing every bar-exam training record against every benchmark question in the same language, no pair matches exactly. The largest character 5-gram Jaccard score is 0.463. A different surface metric, SequenceMatcher, reaches 0.717 for a Limitation Act schedule question paired with benchmark item 2022-086. They use different wording but test the same limitation period. We therefore disclose substantive overlap rather than treating the absence of a verbatim match as proof of separation.

Failure modes engineered out. Five recurring generation failures were identified and countered in the prompts rather than filtered afterwards. Models paraphrased or summarised statutory language instead of quoting it, so the prompts forbid altering legal text and require verbatim copying. Long sections with many subclauses were partially covered, so the prompts require every clause to be enumerated before an answer is produced. Repeated questions were generated for the same provision, so the prompts require type diversity from a fixed vocabulary and forbid using the same type twice in a row. Sections were occasionally drawn from a different act than the one being processed, so the prompts require the act name to be stated first and the citation to be the final key. Format compliance decayed over long runs, so each prompt ends with a pre-output checklist that the model must satisfy.

Human review. Every candidate record was checked by the authors against the source law. The protocol had three parts: fact-checking each answer against the provision it cites, verifying that quoted statutory text was copied rather than summarised, and deleting or regenerating records that were repetitive, drew on more than one act, or omitted a material clause. Of 4,125 candidates, 2,165 were retained. Review ran alongside generation, so no per-record log was kept. The earliest preserved consolidated file holds 3,425 records; the released set is a strict subset of it, verified by identifier and by question text (`prefilter_pool.py`).

Generation prompt, single-hop extraction, English. Reproduced in the form used, with the example entry abbreviated. The other five prompts follow the same skeleton with category-specific constraint blocks, and all six are in the release.

```
Role:  Expert Legal Educator and Subject Expert.
Task:  Analyze the provided JSON legal dataset and transform each section of the Act into a
structured assessment and reference entry.

[SFT Data Quality Context]
You are generating supervised fine-tuning (SFT) data for smaller language models (0.6B-8B
parameters).  Your outputs are the training signal.  Smaller models learn by pattern
recognition:  they need structural consistency.  Length variance and format drift degrade
SFT quality directly.

[Jurisdiction & Corpus]
```

```
Legal Corpus: Use only the Act names, section numbers, and legislative addresses provided
in the dataset. Never invent a Section number if the provision uses a different label.
[Format & Schema]
Output must be a strictly valid JSON array of objects. Schema per object: Section
Number; Entry_ID; Question (plain layperson language, no legal jargon in the stem);
Subsection/Clause; Section Text (full, verbatim); Answer; Type; Difficulty; Keywords; Cited
Acts and Sections (must be the last key).
[Field Rules: The Answer Field]
Strict Opening: Must begin with "Under [Act Name]," followed immediately by the precise
legislative address. Constraint: one to two sentences maximum, 25-45 words.
[Constraint Guidelines]
Verbatim Integrity: Copy "Section Text" exactly. Do not truncate, paraphrase, or
reconstruct from memory. Client-Speak Rule: questions must sound like a client describing
a problem. Scenario Freshness: each scenario must use a distinct factual context; do not
reuse character names across consecutive entries. Multi-Clause Sections: generate one
question targeting the most significant clause and note it in the "Subsection/Clause" field.
[Pre-Output Quality Check]
[ ] Is Section Text copied verbatim? [ ] Does the Answer start with "Under [Act Name]," and
stay under 45 words? [ ] Is the Question written in plain, non-legal language? [ ] Is the
JSON strictly valid with "Section Number" first and "Cited Acts" last?
```

The Bangla prompts are the same skeleton with the instruction text and the answer-field rules in Bangla, and with an additional type-selection procedure that draws from a fixed question-type vocabulary and forbids repeating a type consecutively. They are released as files rather than reproduced here because this template cannot typeset Bangla script.

Generation prompt, advanced selection, English. The constraint block that distinguishes this category from single-hop extraction. The rest of the prompt shares the skeleton above.

```
Possible Sections: A list of exactly 5 objects, each containing Section Number and Full
Text (the correct section + 4 distractors). Full text must be copied verbatim from the law
dataset.
The governing section's position - first, second, third, fourth, or fifth - must vary
randomly across entries, with approximately equal frequency. Never default to always
placing it first.
Multi-Clause Sections: Target the specific clause that governs the scenario and note it in
the Subsection/Clause field.
```

Generation prompt, bar-exam category, English. This category is pinned to a supplied reference file rather than composing freely, which is why its prompt speaks of copying rather than writing. The reference collection holds bar-examination-style questions from examinations other than the two we evaluate on; the model's work here is locating the governing provision and the four distractors, not inventing the stem. We reproduce the prompt as run, and note that we can vouch for what was supplied to it, not for how that reference collection was itself assembled.

```
Task: You are provided bar examination questions from the supplied JSON dataset, each
with multiple-choice options (A, B, C, D) and a known correct answer. For each question,
identify the one governing legal provision from the supplied law dataset, produce a full
Answer citing the relevant section, and state the correct option letter as the final element
of the Answer field.
Question: The verbatim question_text from the bar exam JSON, exactly as written. Do not
rephrase or correct.
Bar_Exam_Options: The options object from the bar exam JSON - keys A, B, C, D with their
verbatim text. Copy exactly.
Cancelled: Boolean. Set to true if correct_answer in the bar exam JSON is "Cancelled"
or equivalent. When true, set Relevant Section to "N/A" and Answer to "This question was
```

```
cancelled in the original exam."
Legal Corpus:  Use only the Act names, section numbers, and legislative addresses that
appear in the provided law dataset.  Do not import terminology or citation formats from
other jurisdictions.
```

The jurisdiction constraint in the last block exists because closed-book prompting produced confident citations to Indian statute in place of the Bangladeshi equivalent, which is the failure the corpus and the supplied context are meant to remove.

C Training Details

Frozen prompt. Every system, in both arms and in training, receives the same system message:

```
You answer Bangladesh law questions using the supplied legal context when it is relevant.
If the question contains answer options, return exactly one line and nothing else in this
form:  FINAL_ANSWER: (A|B|C|D). For Bangla option labels, A, B, C, and D map to the four
Bangla option glyphs.  If the question has no answer options, answer the question directly.
```

The released `bdlex_core.py` carries the Bangla glyphs verbatim in place of that clause, together with the option-normalization map that accepts Latin, numeric, and Bangla labels. The user turn is `LEGAL_CONTEXT:` followed by the context block, then `QUESTION:` followed by the question, with the literal string `[NO CONTEXT PROVIDED]` substituted under the `none` condition. Cancelled items are detected by marker matching on the answer field in either language and dropped before scoring.

Training recipe. Held constant across all six models: sequence length 1,536, per-device batch 4, gradient accumulation 4, effective batch 16, LoRA rank 16 and alpha 16 with no dropout on all linear language-model projections (`q,k,v,o,gate,up,down`), 8-bit AdamW, weight decay 0.01, cosine schedule, warmup ratio 0.06, gradient clipping 1.0, gradient checkpointing on, packing off, at most two epochs, early stopping with patience of one evaluation on validation loss. Embedding and output-head modules are untrained because no tokens are added, and training runs through Unsloth (Han et al., 2023). Two epochs at effective batch 16 is 216 optimizer steps. Records are split by normalized (act, section) group with a fixed splitter seed, giving 1,728 training, 220 validation, and 217 internal-test records with no statutory group in more than one split. Each model was trained three times, with seeds 17, 42, and 73, on a single 16 GB T4.

Early stopping fired for Qwen3.5-2B seed 42 and for all three Qwen3.5-4B seeds, each halting at step 162 with the best checkpoint retained; every other run completed 216 steps. Final training and validation loss were 0.556/0.568 for Llama-3.2-1B, 0.432/0.460 for Llama-3.2-3B, 0.931/0.903 for Qwen3.5-0.8B, 0.796/0.768 for Qwen3.5-2B, 0.691/0.647 for Qwen3.5-4B, and 0.132/0.875 for Gemma-4-E2B, the only model whose validation loss finished far above its training loss.

Learning-rate pilots. Each pilot ran one epoch on the same deterministic 512-row training subset with seed 17, scoring only grouped validation loss over the pre-specified grid $\{5e{-}5, 1e{-}4, 2e{-}4\}$, with ties inside 0.005 broken toward the lower rate. All six selected 2e−4 with validation loss falling monotonically across the grid; for Llama-3.2-1B the three values were 1.0002, 0.8490, and 0.7426. The pre-specified rank-escalation contingency, which would have repeated a pilot at rank 32 had validation loss failed to improve by 0.02, did not fire for any model.

Gemma-4-E2B corrections. The reported run is the second of seven attempts, and the first was wrong in three ways that we record because each could have produced the reversal on its own. The LoRA target pattern was unconstrained, so half the adapter landed on vision and audio components. The training template concatenated prompt and target, so loss was taken

over the instruction as well as the answer, and generation emitted an empty reasoning block that training had never seen. Shuffled-control candidates were drawn without restricting language or length. The repair restricts the adapter to 245 language-model projection modules, 490 LoRA tensors, with no multimodal parameter trainable, replaces concatenation with explicit prompt-then-target masking, and draws shuffled candidates from the question's own language pool. The reported run satisfies all three.

The pre-repair run also degraded, from 26.4% to a 23.7% seed mean. We do not present that as a replication. It predates the audit, its 1,592-row pooled denominator is not the per-condition denominator of Table 2, and its per-item outputs were not retained as a separate benchmark directory, so we cannot recompute it under the reported protocol. What it supports is only that the direction was present before the three defects were fixed.

Within the reported run the seed-42 training crashed near the end on a learning-rate assertion and was benchmarked from un-converged weights, so we retrained that seed to its full 216 steps and benchmarked again.

The two benchmark executions can be compared file by file, and the comparison is the reason we do not treat the degradation as an artifact of that crash. The reference, seed-17, and seed-73 result files are byte-identical across the two runs, the seed-42 file differs, and every cell of the strict-consistency grid, including the seed-42 column, is unchanged. Replacing an un-converged adapter with a fully converged one moved individual predictions without moving a single rotation-invariant score. We note explicitly that this makes the two executions one experiment with a repaired seed rather than two independent replications of the model.

Adapter scope, before and after the repair. The two adapters can be compared directly, and the leak is visible in their weights rather than only in our description of it. The pre-correction target pattern matched `vision`, `audio_tower`, `embed_vision.embedding_projection`, and `vision_tower.patch_embedder` alongside the language projections. It adapted **494 modules, of which 249 were multimodal**. The corrected pattern matches only language and text branches and adapts **245 modules, none multimodal**, which is the 490 LoRA tensors reported above. Half of the first adapter's capacity was spent outside the language model.

Training runs on this model. Gemma-4-E2B took seven training attempts, more than any other model. Table 3 lists the ones that recorded metrics. Only the second is reported in this paper; the rest are training diagnostics, none was benchmarked, and none contributes a number to any other table.

Run	LR	step	train	eval	gap
v20 reported	2e−4	216	0.132	0.876	+0.743
v4 gap stop	2e−4	54	0.129	0.990	+0.861
v9 gap stop	1e−5	30	0.571	3.967	+3.396

Table 3: Gemma-4-E2B training runs that recorded metrics. Seed means over three seeds. Only the first is benchmarked or reported elsewhere. Recomputed by `gemma_run_ablation.py`.

Three further attempts left no metrics: the first, superseded run, whose adapter config carries the multimodal leak described above; one whose notebook never submitted; and one that crashed on a learning-rate assertion. A seventh run trained at 1e−5 under a relative-gap rule and reached step 216 for two seeds before we cancelled it with the gap still near 1.4. Only its notebook was archived, so we report its existence and not its losses.

What the diagnostics show. The 2e−4 runs over-adapt quickly: the gap reaches 0.86 within 54 steps, a quarter of training. The 1e−5 run is more interesting for what it revealed about the instrument. It halted at step 30, but the gap was *shrinking* across those steps, from +4.20 to +3.85 to +3.45. Gemma begins at a validation loss of 4.81 before any training, so an absolute-gap threshold is above its trigger from the first evaluation and fires however well training is

going. The fault was in the stopping rule, not the learning rate. A relative criterion, halting only when the gap grows from its own minimum, is the correct instrument.

The consequence for this paper stays narrow. We know the default rate over-adapts this model quickly, and we do not know what a gentler rate does to its benchmark score, because no lower-rate adapter was trained to completion and scored under the frozen protocol. We did not substitute a rescued recipe for one model, since that would compare a tuned system against five untuned ones.

D Fine-tuning Set Composition

Act / schedule	**S-hop** EN/BN	**Adv.** EN/BN	**Bar** EN/BN	**Tot.**
Penal Code	90/127	111/84	39/39	490
Evidence Act	66/94	109/75	40/33	417
CrPC	66/104	28/108	59/40	405
Limitation Act	36/42	38/43	41/40	240
Specific Relief	17/66	6/68	40/20	217
CPC	27/60	13/29	29/39	197
CPC Sch. I	n/a	n/a	39/40	79
CrPC Sch. II	n/a	n/a	40/20	60
Limitation Sch. I	n/a	n/a	20/40	60
Total	302/493	305/407	347/311	**2,165**

Table 4: Fine-tuning set by act, category, and language (English/Bangla), recomputed from the released file. Acts are named in full in the text. Schedules are listed separately rather than folded into their parent acts; only the bar-exam category draws on them.

E Difference in Differences by Retrieval Condition

Table 2 differences against retrieval defined as the mean of the BM25 and dense changes within a seed, which is the pair of conditions the pre-specified test family covers. Table 5 gives each condition on its own. The two do not always agree: for Gemma-4-E2B in Bangla the contrast against no context is +0.2 points under BM25 and −3.9 under dense, and for Qwen3.5-4B in English it is −4.5 under BM25 and +0.8 under dense. Reporting either condition alone would therefore have supported a different count of models showing context-specific gains, which is the reason we average.

	vs. none			**vs. irrelevant**		
Model	BM25	dense	mean	BM25	dense	mean
Bangla						
Llama-3.2-1B	+3.0	+6.0	+4.5	+2.5	+5.5	+4.0
Llama-3.2-3B	+5.9	+10.4	+8.1	+0.7	+5.2	+2.9
Qwen3.5-0.8B	+5.2	+4.0	+4.6	+4.4	+3.2	+3.8
Qwen3.5-2B	−2.2	−0.2	−1.2	+1.0	+3.0	+2.0
Qwen3.5-4B	+1.3	+5.7	+3.5	+3.7	+8.0	+5.9
Gemma-4-E2B	+0.2	−3.9	−1.8	+3.9	−0.2	+1.8
English						
Llama-3.2-1B	+7.0	+12.4	+9.7	+7.7	+13.1	+10.4
Llama-3.2-3B	+3.9	+0.2	+2.0	+1.7	−2.0	−0.2
Qwen3.5-0.8B	+6.9	+8.5	+7.7	+4.4	+6.0	+5.2
Qwen3.5-2B	−4.2	+0.5	−1.8	−0.3	+4.4	+2.0
Qwen3.5-4B	−4.5	+0.8	−1.8	−1.8	+3.5	+0.8
Gemma-4-E2B	−1.8	−4.5	−3.2	+2.2	−0.5	+0.8

Table 5: Difference in differences by retrieval condition, in points. Values are seed means. **mean** averages the BM25 and dense changes within each seed and is the column reported in Table 2.

F Additional Results

Model	ref.	FT mean	change
Llama-3.2-1B	57.3	26.8	−30.5
Llama-3.2-3B	34.7	15.7	−19.0
Qwen3.5-0.8B	29.6	5.9	−23.7
Qwen3.5-2B	23.1	15.2	−7.9
Qwen3.5-4B	31.7	14.6	−17.1
Gemma-4-E2B	13.6	18.4	+4.8

Table 6: Position-bias spread on Bangla dense rows, in points. The spread is the range of plain accuracy across the four positions the gold option can occupy. The fine-tuned column takes the seed mean per position before the spread, so it is not the mean of per-seed spreads.

Why rotation, and not seed agreement. We considered requiring an item to be answered correctly by all three training seeds and rejected it before scoring. The reference is a single deterministic run with no seed axis, so a three-way agreement requirement would apply only to the fine-tuned arm and would penalise it for a source of variation the comparison arm does not have. Option rotation is the axis both arms possess, so it is the axis the strict metric uses.

	Bangla			English		
System	rot0	strict	drop	rot0	strict	drop
Llama-3.2-1B	34.7	2.0	32.7	47.2	15.6	31.7
Llama-3.2-3B	42.7	14.1	28.6	61.3	43.7	17.6
Qwen3.5-0.8B	44.2	19.6	24.6	51.8	31.2	20.6
Qwen3.5-2B	45.2	23.1	22.1	56.8	38.7	18.1
Qwen3.5-4B	55.8	28.6	27.1	67.8	50.8	17.1
Gemma-4-E2B	54.8	37.7	17.1	61.8	49.7	12.1

Table 7: Effect of option rotation on reference systems, dense condition. Here **rot0** is plain accuracy at one option order, **strict** the percentage correct under all four rotations, and **drop** the difference. Recomputed by `rotation_stability.py`.

Holm-significant counts within the pre-specified 12-test family per model are 12/12 for Llama-3.2-1B, 3/12 for Llama-3.2-3B, 7/12 for Qwen3.5-0.8B, 0/12 for Qwen3.5-2B, 2/12 for Qwen3.5-4B, and 0/12 for Gemma-4-E2B. Per-seed differences, exact McNemar p values, Holm-adjusted p values, and 10,000-draw item-clustered bootstrap intervals for all 72 comparisons are in the released analysis JSON rather than reproduced here. Those intervals are on the fine-tuning difference under one retriever at one seed. The R−N and R−I intervals in Table 2 are computed separately from the same per-item rows by `bootstrap_did.py`, which resamples items and checks each point estimate against the table before reporting a width. Table 8 gives the intervals underlying Table 1: the law-effect and FT contrasts come from `bootstrap_gold.py`, following the same resampling pattern on the 150 control items and the 199 examination items with both language readings drawn together. The provision and same-act decomposition intervals come from `bootstrap_gold_withheld.py`, which applies the same resampling to the withheld-provision run.

Sparse cells. Strict consistency puts several cells near the floor, where a percentage hides how few items are involved. Table 9 gives the item counts for every cell of Table 2 at or below 5%, recomputed from the same per-item rows by `sparse_cell_counts.py`.

The parser diagnostic across seeds. The seed-42 column is not a special case. Recomputing the same three readings for every seed with `recompute_lenient.py`, the Gemma-4-E2B sign disagreement holds throughout: the parser reads +11.3, +10.1 and +10.8 for seeds 17, 42 and 73, with lenient parsing identical to strict at each seed, while constrained scoring reads −1.5,

Model	law effect	FT−Ref, none	FT−Ref, law	control	exam R−N
Llama-3.2-1B	−1.3	+16.2	+14.4	−1.8	+7.1
	[−7.3, +4.7]	[+10.4, +22.4]	[+7.6, +21.3]	[−9.6, +5.8]	[+4.1, +10.2]
Qwen3.5-0.8B	+18.7	+13.3	+14.4	+1.1	+6.2
	[+11.3, +26.0]	[+7.8, +19.3]	[+8.0, +21.1]	[−7.6, +9.8]	[+3.2, +9.0]
Llama-3.2-3B	+17.3	+3.1	+2.2	−0.9	+5.1
	[+10.0, +25.3]	[−2.2, +8.4]	[−4.7, +8.9]	[−8.4, +6.7]	[+1.2, +9.1]
Qwen3.5-2B	+19.3	+2.7	+1.8	−0.9	−1.5
	[+12.0, +27.3]	[−2.0, +7.3]	[−3.8, +7.3]	[−8.2, +6.4]	[−5.2, +2.1]
Gemma-4-E2B	+14.7	−11.1	−6.0	+5.1	−2.5
	[+6.7, +22.7]	[−17.6, −4.9]	[−11.3, −0.7]	[−2.2, +12.4]	[−5.6, +0.6]
Qwen3.5-4B	+17.3	−7.6	−2.9	+4.7	+0.8
	[+10.7, +24.0]	[−13.6, −1.8]	[−6.9, +1.1]	[−2.0, +11.1]	[−2.5, +4.1]

Table 8: Differences and intervals underlying Table 1. **control**: (FT−Ref, law) minus (FT−Ref, none). Control columns: $n = 150$; **exam R−N**: $n = 199$.

Model	condition	sys	k/199
Bangla			
Llama-3.2-1B	none	Ref	1 (0.5%)
Llama-3.2-1B	BM25	Ref	1 (0.5%)
Llama-3.2-1B	irrelevant	Ref	1 (0.5%)
Llama-3.2-1B	dense	Ref	4 (2.0%)
Qwen3.5-2B	none	Ref	5 (2.5%)
Qwen3.5-0.8B	none	Ref	7 (3.5%)
Qwen3.5-0.8B	none	FT	8 (4.0%)
Llama-3.2-3B	irrelevant	Ref	8 (4.0%)
English			
Llama-3.2-1B	irrelevant	Ref	8 (4.0%)

Table 9: Item counts behind every Table 2 cell at or below 5%, out of 199 items. k counts items answered correctly under all four rotations. **FT** counts are the mean over the three seeds. The remaining 87 cells sit above 5%.

	reference				seed-42 gain		
Model	fmt.	strict	lenient	logit	strict	lenient	logit
Qwen3.5-2B	5.8	4.0	33.2	51.0	+50.0	+20.9	+3.0
Llama-3.2-1B	6.0	2.8	32.2	41.0	+39.7	+10.3	+1.8
Qwen3.5-0.8B	80.7	41.7	45.0	48.0	+9.8	+6.5	+3.5
Gemma-4-E2B	80.9	45.5	45.5	58.3	+10.1	+10.1	−2.5
Llama-3.2-3B	81.7	45.5	54.0	52.0	+9.3	+0.8	+2.8
Qwen3.5-4B	98.7	62.6	63.1	61.8	−0.8	−1.3	+0.0

Table 10: One comparison read four ways, ordered by reference format compliance. Rows are the 398 dense-condition rows at rotation 0. **fmt.** is the percentage of reference generations satisfying the answer grammar, **strict** requires one matching line, **lenient** accepts a letter anywhere, and **logit** is the constrained reading. Values are plain rotation-0 accuracies against a most-frequent-label baseline of 29.1%, so they are not comparable to the four-rotation strict-consistency values in Tables 1 and 2. Lenient parsing removes much of the inflation at the two least compliant models but not all of it, and it leaves the Gemma-4-E2B sign disagreement unchanged.

−2.5 and −1.8. The Qwen3.5-2B inflation is likewise stable, at +52.8, +50.0 and +50.8 strict against +5.5, +3.0 and +3.8 constrained.

Quoted-text sensitivity. The control's most-quoted-option heuristic scores 56/150 (37.3%). As a reviewer-requested post-hoc check, we remove every tie and keep the 92 items where a distractor has strictly greater 5-gram coverage of the supplied law than the correct option. Reference law effects remain +12.0, +19.6, +16.3, −2.2, +22.8, and +13.0 points for Qwen3.5-0.8B, Qwen3.5-2B, Qwen3.5-4B, Llama-3.2-1B, Llama-3.2-3B, and Gemma-4-E2B, respectively.

`handbuilt_gold_results.py` recomputes the subset from recorded contexts, options, and per-item outputs. This analysis was not pre-specified.

The length-matched irrelevant control. This is the only post-hoc evaluation condition. It replaced the frozen protocol's shorter shuffled control, fills the retrieval-length budget with sections from the question's own language pool that are excluded from its BM25 and dense hits: median 2,000 characters in both languages, against 2,002 for both retrieval conditions. Because the control matches retrieval length, a contrast against it separates relevance from the amount of supplied text, so R−I and R−N are both reported in the main text.

G Translation Audit

Planted error	n	judge A	judge B
key option swapped	8	8	2
non-key option replaced	8	8	1
legally inert edit (false positive)	6	1	0

Table 11: Judge sensitivity on mechanically planted corruptions with known ground truth. Entries count items rated major or critical; for inert edits a flag is a false positive. Judge A's 1 of 6 there has a 95% Wilson interval reaching 56%.

	all items			by year	
Model	199	−8	−25	2022	2023
Llama-3.2-1B	+9.7	+10.5	+10.2	+7.6	+11.8
Llama-3.2-3B	+2.0	+1.3	+4.1	+0.3	+3.7
Qwen3.5-0.8B	+7.7	+7.9	+9.4	+5.6	+9.8
Qwen3.5-2B	−1.8	−1.7	−1.6	−2.2	−1.5
Qwen3.5-4B	−1.8	−1.6	−2.2	−4.9	+1.2
Gemma-4-E2B	−3.2	−3.3	−3.4	−3.0	−3.3
items flagged by judge A				29/99	0/100

Table 12: English difference in differences on benchmark subsets, in points. The contrast is that of Table 2, recomputed from the raw per-item rows. −8 drops the items judge A called answer-changing and −25 every never-seeded item it flagged, leaving 191 and 174. The last two columns split by examination year. All 29 flags lie in 2022.

Each judge saw the Bangla item and the English item side by side and returned a severity in {none, minor, major, critical}, whether the recorded answer is preserved, whether the option set changed, whether terms of art survived, and a free-text note, at temperature 0. A flag means major or critical. Judge A is `gemini-3.6-flash` and judge B is `gemini-2.5-flash-lite`, and the translation came from a different vendor's model (Section G), so no judge evaluates its own family's output. The planted substitutions test detection of gross answer-changing errors; they do not measure sensitivity to subtle mistranslation of legal terms.

Restricted to the 177 items that were never seeded, judge A flags 25 (14.1%, 95% CI 9.8 to 20.0) and preserves the recorded answer on 168 of 177 (94.9%); judge B flags 3 of 176 (1.7%, 95% CI 0.6 to 4.9) and preserves the answer on 175 of 176. Counting all 199 items with the seeded ones restored to clean text, A flags 29 and B flags 4 of the 198 for which it returned a verdict. The 8 items A calls answer-changing and the 25 it flags are the two exclusion sets in Table 12.

H Translation Robustness

A translation confound in the English arm has to survive Table 12. It does not. Dropping the 8 items judge A calls answer-changing moves every English difference in differences by at most

0.8 points, and dropping all 25 flagged items by at most 2.1, with no sign flip. Translation noise would also inflate the year holding all 29 flags, yet 2022 is the smaller year for five of six models.

I Correctness or Position Robustness

Strict consistency requires the correct answer under all four option rotations, so supplied law can raise it two ways: by making more answers right, or by making an already preferred answer survive rotation. Table 13 separates them, reading the control at rotation zero and under all four in the format of Table 7.

	law		no law		law effect	
Reference	rot0	strict	rot0	strict	rot0	strict
Llama-3.2-1B	63.3	27.3	56.7	28.7	+6.7	−1.3
Qwen3.5-0.8B	74.7	57.3	58.0	38.7	+16.7	+18.7
Llama-3.2-3B	84.7	68.0	73.3	50.7	+11.3	+17.3
Qwen3.5-2B	88.0	76.7	76.7	57.3	+11.3	+19.3
Gemma-4-E2B	92.7	84.7	79.3	70.0	+13.3	+14.7
Qwen3.5-4B	98.7	94.0	86.7	76.7	+12.0	+17.3

Table 13: The control's law effect read two ways, reference systems only, in percentage points. **rot0** is plain accuracy at one option order and **strict** is the percentage correct under all four. Rows are ordered by the rot0 law score. Recomputed by `control_rotation_decomposition.py`.

Two things follow. The law effect is positive at rotation zero for every model, Llama-3.2-1B included, so all six use the provision when the option order is held fixed. And for the five responsive models the strict effect is the larger of the two, by 1.3 to 8.0 points, so part of what strict scoring credits is a reduction in position brittleness rather than a new correct answer. The effect is present under both readings for those five, so it is not only robustness, but the strict figure should not be read as pure correctness either.

Llama-3.2-1B is the case where the two readings disagree in sign: +6.7 at rotation zero against −1.3 under strict scoring. Supplied law moves its preferred answer without making that answer stable across orders. This is the measurement behind the option-label prior described in Section 6.2, and it narrows the claim: the model is not unable to use the provision, it is unable to hold the answer when the options move.

J Reproducibility

Availability. The corpus, the QA dataset, the benchmark, the adapters and the analysis code are public under permissive terms, and are described in enough detail throughout this appendix that a reader can judge what exists without following a link. The dataset is publicly available at `https://huggingface.co/datasets/momahadi/bangladesh-legal-qa-dataset` and the code and artifacts at `https://github.com/m-mahadi/bangladesh-legal-qa`; together they expose the contract, corpus, QA dataset, benchmark, translation audit, per-item outputs across all eight MCQ executions, free-form generations, the Gemma training-run registry, the `MANIFEST.csv`, and every analysis script. Licensing is stated in the Ethics Statement.

From the repository root, the command `python 07-analysis/scripts/reproduce.py` runs the 27 analyses that regenerate every number in this paper in about a minute, without a GPU or network access: hard gates fail loudly on any mismatch, soft checks print the value the paper quotes for manual comparison, and `python 07-analysis/scripts/reproduce.py --list` prints the claim map tying each script to the claim and the folder it regenerates.

The release contains the structured bilingual corpus, the 2,165-record QA set and the pre-review pool it came from, the group-disjoint training split with its manifest and group-assignment digest, the frozen retrieval bundle, the experiment contract, every adapter, every per-item benchmark output, the raw free-form generations, the Gemma training runs with their notebooks, the run manifests, and the analysis scripts that regenerate the reported tables and diagnostics from

retained outputs in one command. This is numerical reproducibility, not a claim that every archived run has complete provenance; the two known integrity gaps are stated below. Each benchmark run directory carries a manifest recording the model revision, prompt hash, contract hash, scorer identifier, package versions, hardware, peak memory, and a SHA-256 for each result file. The release holds the artifacts this paper reports on and nothing else: earlier exploratory cycles, including fine-tuning runs on models outside the six evaluated here, are excluded, and the frozen contract fixes the model matrix independently of them.

Checking the freeze. One unedited contract and one unedited prompt governed every reported run, and that is checkable from the release without trusting us. Hash the released contract file and compare it against the `contract_sha256` field each run wrote into its own manifest at execution time. All eight benchmark runs, covering all six models and both Gemma executions, record `419b73b8b1b33e0a8298b9bd15801cad2c87bd9b07ea62af7ac5ee9eb3ef8f02`, and the released contract hashes to that value. The same manifests carry one prompt digest, `ec2c0196c3db4879612433e528f34d1c2050ebb13656a4535436741c12629d18`, across every run. A contract edited after a result was seen would not reproduce either digest. The digests do not by themselves establish the calendar order in which the runs happened, for which the repository commit history is the record.

The five non-Gemma models were additionally re-audited by an independent script that reparsed all 40 result files, confirmed every recorded hash, and verified 199 item identifiers, two languages, four conditions, four rotations, no duplicate composite keys, no cancelled item, and a key set identical to the matched reference across all 127,360 constrained rows before recomputing every statistic reported here.

Two gaps are recorded rather than hidden. The archived Qwen3.5-0.8B and Qwen3.5-2B benchmark directories lack the `RUN_COMPLETE` marker file, although both carry terminal manifests with end times and all eight result hashes verified. And for the corrected Gemma-4-E2B adapters the recomputed local directory-tree hashes do not reproduce the tree hashes recorded in the training manifests, even though the file inventory matches and every safetensors file is complete; the individual weight-file SHA-256 values are therefore the durable provenance for that model, and the benchmark outputs produced with those adapters are manifest-hash verified.

K Expert Review of the Supplied-Law Control

After all experiments were complete, two independent reviewers assessed every control item against its governing provision. Neither reviewer saw model outputs, experimental results, or the other reviewer's ratings. Table 14 summarises their judgments.

Reviewer	Good	Acceptable	Needs work
Law graduate	120	28	2
Law masters student	148	–	2

Table 14: Independent post-experiment review of the 150 supplied-law control items. The 28 acceptable ratings cite minor stylistic concerns (option length balance, distractor difficulty, terminology mixing). The four items flagged as needing work fall on different items across reviewers (no overlap). Both reviewers confirmed that no answer is incorrect given its cited provision. No items were changed.