

Do Small Models Use the Law You Give Them? Context-Injected Fine-Tuning for Legal QA in Bangladesh

Moniruzzaman Mahadi¹, Abrar Mohammed Tanzim Alam¹, Sayma Siddika Monalisa¹
Mir Mohammad Asif Abdullah¹, Swakkhar Shatabda², Md Adnan Arefeen¹

¹Department of Electrical and Computer Engineering, North South University
Dhaka, Bangladesh

²Department of Computer Science and Engineering, BRAC University
Dhaka, Bangladesh

Correspondence: momahadi9664@gmail.com

Abstract

A small language model can receive the governing statutory provision and still answer incorrectly. We test whether fine-tuning on examples containing relevant law improves later use of retrieved law. We curate 2,165 bilingual QA records from six Bangladeshi acts and three schedules, then fine-tune Qwen3.5 at 0.8B, 2B, and 4B. Evaluation uses the 2022 and 2023 Bangladesh Bar Council exams in Bangla and machine-translated English, with no retrieval, BM25, or FAISS, scored by strict consistency over three seeded runs. At 0.8B, fine-tuning raises the 2022 English FAISS score from 2 to 34 of 100. Gains at 0.8B and 2B survive paired testing, but the 4B model has no detectable net gain: Bangla improves while several English conditions regress. Fine-tuning also reduces answers that drift from Bangla into mostly English from 44.0–53.2% to 0.2–0.7%, with adjusted $p < .001$ at every scale. Retrieval quality is therefore not the only bottleneck. Small bilingual legal models also differ in how they use supplied law and whether they answer in the requested language. The dataset is publicly available at <https://huggingface.co/datasets/momahadi/bangladesh-legal-qa-dataset>.

1 Introduction

Article 27 of the Constitution of Bangladesh guarantees equality before the law (Government of the People’s Republic of Bangladesh, 1972). Yet legal help remains out of reach for much of the population. Analyses of the Bangladeshi justice system describe several million pending cases handled by roughly two thousand judges, prolonged litigation, unregulated lawyer fees, and narrow eligibility for public legal aid (Yesmin, 2025; Ahmed, 2021; Jaan, 2023; Tahura and Alam, 2025; Islam

et al., 2024; Akter, 2017). Language technology will not fix the court system, but it can lower the cost of reaching accurate legal information, provided its answers stay grounded in the governing statute.

Much recent legal NLP pursues that grounding through retrieval: a better index puts the right provision into the model’s context. Small models strain that assumption. In our experiments, they often failed with the governing section already in the prompt, and Bangla was worse. Bengali remains low-resource for current model families (Bhowmik et al., 2025). Bangladeshi legal language also combines Persian-influenced and colonial-era terms with standard Bengali (Wasi et al., 2025). Even frontier models have struggled with authentic Bangladeshi legal queries (Aftahee et al., 2025).

We focus on what happens after retrieval. Does supervised fine-tuning on legal QA pairs whose prompts already contain the relevant statutory text improve a small model’s use of supplied context at inference time? We call this ability *context utilization* and measure it as the change in multiple-choice accuracy with and without retrieved statutory context, before and after fine-tuning. Figure 1 summarizes the study. Our benchmark is the 2022 and 2023 Bangladesh Bar Council MCQ examinations, taken in the original Bangla and in machine-translated English. MINA’s best reported cells on these examinations are 75.6–77.0%, obtained with the proprietary Gemini-2.5-Flash backend under two-step RAG or tool use; its small open-weight backends score far lower (Wasi et al., 2025). Those five-run averages are not directly comparable to our strict-consistency score. We hold the retrieval machinery fixed and vary

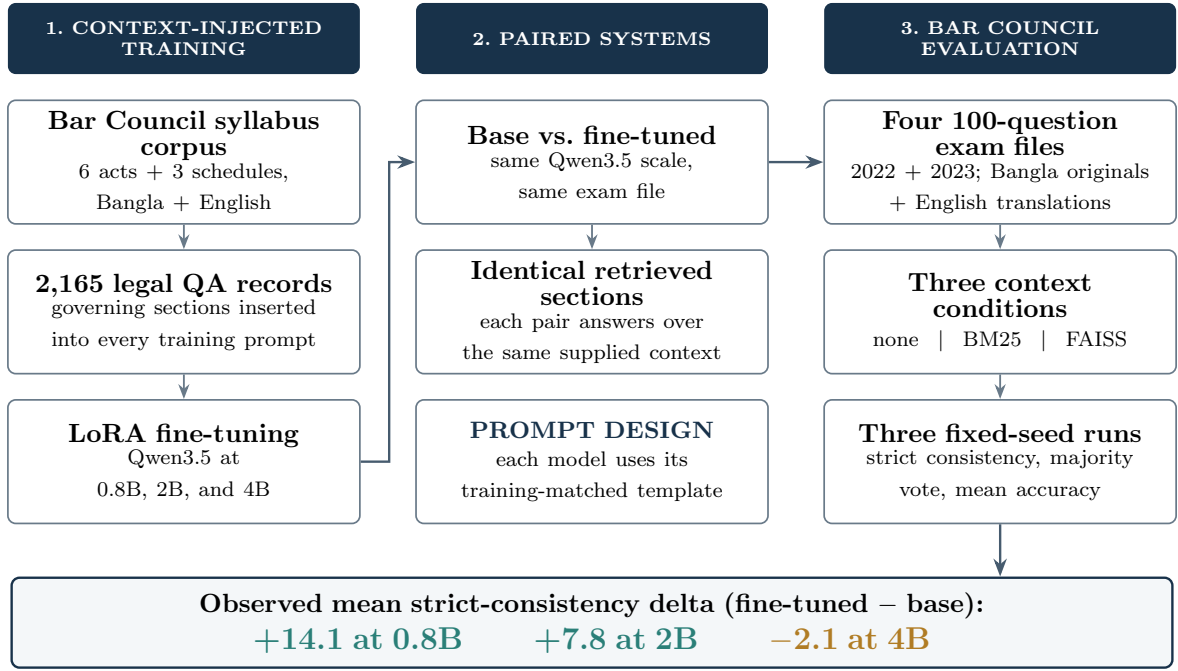


Figure 1: Study design. Context-injected supervision produces fine-tuned variants of Qwen3.5 at three scales. Base and fine-tuned systems are compared on four Bangladesh Bar Council exam files under no retrieval, BM25, and FAISS. Mean strict-consistency deltas favor fine-tuning at 0.8B and 2B but reverse at 4B.

the model: base and fine-tuned Qwen3.5 at 0.8B, 2B, and 4B parameters, each under no retrieval, BM25, and FAISS dense retrieval, three fixed-seed runs per condition. One run of GPT OSS 120B gives an informal ceiling; it is not a replicated baseline.

Our contributions are:

1. a structured, hierarchy-preserving bilingual corpus of the six Bangladeshi acts and three schedules that dominate the Bar Council syllabus (Section 3);
2. a curated bilingual legal QA dataset of 2,165 records in three difficulty formats, with the statutory context relevant to each answer injected into the training prompt (Section 3);
3. a paired comparison of base and context-injected fine-tuned models across scales, languages, and retrieval conditions, reported under strict-consistency, majority-vote, and mean-accuracy scoring, with the regressions made explicit (Sections 5–6).

The effect depends strongly on model size. The fine-tuned models posted large gains at 0.8B and 2B, and retrieval helped them much more than their base counterparts. At 4B, Bangla mostly improved, while several English conditions regressed. Fine-tuning also removed a

large language-drift failure: base models answered Bangla questions mostly in English in 53.2%, 44.0%, and 44.7% of visible outputs at 0.8B, 2B, and 4B. Fine-tuning reduced those rates to 0.7%, 0.2%, and 0.7%, respectively, with adjusted $p < .001$ at every scale. These experiments show gaps in both context utilization and answer-language control across model scale.

2 Related Work

Low-resource legal QA for Bangladesh. Wasi et al. (2024) explored LLM-based legal assistance for Bangladesh by fine-tuning a GPT-2 variant on an English corpus of Bangladeshi statutes. The authors describe a proof of concept with substantial gaps in accuracy and safety. Its English-only setup does not reflect bilingual legal practice. The same group’s MINA (Wasi et al., 2025) is a multilingual legal-assistant system evaluated with several model backends and retrieval configurations on the 2022 and 2023 Bar Council examinations, with law professors reviewing outputs. Its highest table values, 75.6–77.0%, are five-run averages from the proprietary Gemini-2.5-Flash backend under two-step RAG or tool use; they are not directly comparable to our strict-consistency score. The small open-weight models in the same evaluation score far lower. MINA evaluates a

complete assistant across backends and retrieves over the full body of Bangladeshi statute; we instead isolate how small Qwen3.5 systems behave under controlled retrieval, searching only the six acts and three schedules of the Bar Council syllabus. MINA also documents hallucinated provisions, missed statutory conditions, and civil/criminal confusion, and leaves open whether those failures persist when the relevant text is supplied correctly; our experiments test exactly that setting.

Retrieval-augmented legal NLP. RAG (Lewis et al., 2020) grounds generation in retrieved evidence. Over Bangladesh Police Gazettes, Kabir et al. (2025) built a bilingual RAG framework with relevance checking and query refinement before generation. Pipitone and Alami (2024) introduced LegalBench-RAG to evaluate the retrieval step in isolation. They argue that existing benchmarks take retrieval for granted and that long or noisy context can bury the governing provision. LegalBench-RAG evaluates what enters the prompt; our study starts after that step, fixing retrieval and comparing base against fine-tuned models on identical retrieved context.

Bengali capability and legal datasets. Bhowmik et al. (2025) evaluated ten open LLMs across eight translated Bangla datasets, documenting consistent gaps relative to English, especially for smaller models and for families whose tokenizers over-fragment Bangla. Aftahee et al. (2025) evaluated four frontier LLMs on 250 authentic questions from a Bangladeshi legal-advice community with LLM-as-judge scoring plus licensed lawyers, finding well-structured responses alongside dangerous misinformation such as fabricated citations; expert review keeps exposing failures automated scores miss, which also cautions against over-reading our automated MCQ scores. On the dataset side, IndicLegalQA (Venington and Mishra, 2025) evaluated parameter-efficient fine-tuning on structured Indian legal QA, but did not test whether training with supplied context improves the *use* of context at inference time, the distinction central here.

Benchmarks and evaluation validity. KoBLEX (Lee et al., 2025) grounds open legal QA in statutory provisions and asks for multi-hop reasoning over them; IL-TUR (Joshi et al., 2024) spans a far wider set of Indian legal understanding and reasoning tasks. SaulLM-

7B (Colombo et al., 2024) relies on continued pretraining over a large English legal corpus and does not target Bangla statutory QA.

The gap. Existing work separately improves legal retrieval, documents how far small models fall behind in Bangla, and fine-tunes models on legal QA. Within this reviewed set, we found no evaluation of whether context-injected supervision helps a small Bangla-capable model use statutory text supplied at inference time. We test that by fine-tuning small models on bilingual legal QA whose prompts contain the governing sections, then comparing base and fine-tuned models on identical retrieved context.

3 Corpus and QA Data

We recomputed every count in this section from the released dataset. Figure 2 (Appendix A) traces the detailed data, training, and Bar Council evaluation flow.

3.1 Corpus Scope, Sources, and Structure

The corpus covers the six acts and three schedules that carry the most weight in the Bangladesh Bar Council examination 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, plus CPC Schedule I (Orders and Rules), CrPC Schedule II, and Limitation Act Schedule I. We chose this scope because the enrolment examination is our benchmark; the corpus is not the full body of Bangladeshi statute law. Schedule labels follow the released corpus. Each act is bilingual: English text from the official Bangladesh laws portal (<http://bdlaws.minlaw.gov.bd/>), Bangla text from BGS Technologies’ published act collections.

Statute documents are irregular: act, part, chapter, section, subsection, clause, subclause, plus attached illustrations and explanations. We converted each act to JSON with Google’s Gemini on the Antigravity platform, one custom prompt per act, after regular-expression preprocessing had separated illustrations from identically labelled clauses. We iterated a schema-enforcement prompt (Appendix F) until the files conformed, then inspected them manually. For every section the corpus keeps its act, chapter, number, marginal note, full text, and nested structure.

Act / schedule	S-hop EN/BN	Adv. EN/BN	Bar EN/BN	Tot.
Penal Code 1860	90/127	111/84	39/39	490
Evidence Act 1872	66/94	109/75	40/33	417
CrPC 1898	66/104	28/108	59/40	405
Limitation Act 1908	36/42	38/43	41/40	240
Specific Relief 1877	17/66	6/68	40/20	217
CPC 1908	27/60	13/29	29/39	197
CPC Sch. I	–	–	39/40	79
CrPC Sch. II	–	–	40/20	60
Limitation Sch. I	–	–	20/40	60
Total	302/493	305/407	347/311	2,165

Table 1: Final dataset by act, category (single-hop, advanced selection, bar-exam), and language (English/Bangla). Recomputed from the released 2,165-record dataset; schedule rows listed separately rather than folded into parent acts.

3.2 QA Generation

From the corpus we generated instruction-tuning records in three categories of increasing difficulty, each with a dedicated prompt per language (Appendix F): **single-hop extraction** (answer from one supplied provision), **advanced selection** (five candidate sections, four semantically similar distractors, governing provision at a randomized position), and **bar-exam style** (synthetic scenario MCQs generated from the statutory corpus, using earlier bar-examination questions only as style exemplars, with each item annotated with the governing provision, candidates, and correct option). Generation kept failing in familiar ways: paraphrased statutory text, skipped clauses, repetitive questions, cross-act contamination, format drift. We engineered these out through prompt revision, and the released prompts carry the countermeasures.

3.3 Quality Control and Composition

We reviewed the records ourselves: fact-checking answers against the source law, verifying verbatim quotation, deleting or regenerating defective entries. No per-record review log survives, so we report this as our process, not as an audited exhaustive protocol. Every released record carries a per-record `quality_flag`, and none of the 2,165 is flagged. We report only the released count because the pre-filter generation pool was not preserved.

Table 1 gives the verified composition: 2,165 records, 1,211 Bangla and 954 English; 795 single-hop, 712 advanced-selection, 658 bar-exam-style. Only the bar-exam category draws on the schedules, matching how heavily the en-

rolment examination tests Orders, Rules, and limitation periods. Each record carries the question, verbatim provision text, candidate sections where applicable, answer, citation, type, difficulty, language, and source metadata. How the records are split and trained on is described in Section 4.

3.4 SFT Packaging: Context Injection

We prepared two chat-format JSONL variants of the 2,165 records: a direct-answer variant, and an IRAC variant that wrote out the reasoning before the answer. All fine-tuning and evaluation in this paper use the direct-answer variant. The IRAC variant made each training example much longer, and fine-tuning on those longer inputs did not run to completion on our Kaggle setup, so we set it aside. Context injection is the load-bearing choice. Every training example looks exactly like a retrieval-augmented inference prompt: the system message “You are a Bangladeshi legal assistant. Answer the legal question using the provided law sections.” and a “Retrieved Sections:” block in the user turn. During training the block is filled deterministically from the record, governing section plus distractors; no retriever runs. The model should learn to ground its answers in supplied sections, in the same format a real retriever will fill later.

3.5 Evaluation Benchmark

The benchmark is four 100-question files: the 2022 and 2023 Bar Council MCQ examinations in Bangla, and English versions translated from the originals with Claude Sonnet 4.5 (not independently re-verified by legal translators). The training and evaluation sets stay separate by construction. We wrote the training questions from the Bar Council syllabus statutes, not from the exam files, and we withheld the 2022 and 2023 exams from generation; the bar-exam-style prompts used earlier examinations only as format and style exemplars. We confirmed the separation directly: after Unicode normalization, whitespace collapsing, and punctuation stripping, none of the 400 evaluation questions matches any of the 2,165 training questions exactly. The only thing the two sets share is the underlying law. During retrieval-based evaluation, BM25 and FAISS search those same statutes to supply relevant provisions, which is the intended setup and not exposure to the benchmark questions or answer keys. Released artifacts are mapped in Appendix E.

4 Method

Models. We use the Qwen3.5 series at 0.8B, 2B, and 4B parameters (Qwen Team, 2026), chosen because its Bangla output stayed stable in preliminary use without decoding interventions. Training all three scales under one configuration lets model size vary while data, adapter design, and optimization stay fixed. In closed-book trials these models answered from parametric memory and repeatedly reached for Indian rather than Bangladeshi law, citing the Indian CrPC 1973 in place of the Bangladeshi CrPC 1898. That failure motivated the grounding-first design.

Training-time injection vs. inference-time retrieval. Two separate mechanisms put law into the prompt. At training time, the “Retrieved Sections” block is filled deterministically from the dataset (Section 3.4); no retriever is involved. At inference time, the same block is filled three ways as experimental conditions: left empty, filled by BM25, or filled by FAISS dense retrieval. The hypothesis is that a model fine-tuned on the deterministic version extracts more value from the retrieved version than its base counterpart.

Fine-tuning. We trained LoRA (Hu et al., 2021) adapters via Unsloth (Han and Han, 2024) ($r=32$, $\alpha=64$, dropout 0.05) on all seven linear projections; about 1% of parameters were trainable (12.8M of 0.87B at 0.8B, 21.8M of 2.2B at 2B, 42.5M of 4.6B at 4B). We requested QLoRA-style (Detrmers et al., 2023) 4-bit base-model loading in every training run, but we cannot confirm that the quantization engaged for this architecture: our logs record only a compute-dtype fallback to float32, and we kept no per-layer quantization check or memory figure that would settle it. We therefore report the configuration we set and stop short of claiming whether the base weights ran in 4-bit or full precision during training. Training used AdamW 8-bit, learning rate 2×10^{-4} with cosine decay, weight decay 0.01, 2 epochs, and 2,048-token sequences, with loss on assistant tokens only and thinking mode disabled (full configuration in Appendix A). In practice we trained with an effective batch of 32 over 122 optimizer steps at every scale, and 14 warmup steps, about 11% of the schedule. We had configured a batch of 16 and a 6% warmup, so these are the values the runs actually used, and we report the executed numbers throughout. Our 90/10 split (seed 42)

gives 1,948 training and 217 validation examples with no over-length exclusions, the same at 0.8B, 2B, and 4B. We reused the fine-tuning system prompt verbatim at inference, keeping the training and inference templates matched; fine-tuned models can be sensitive to template changes between the two stages (Lyu et al., 2024). After training, we merged each adapter into the bfloat16 base-model weights and exported the merged model to F16 GGUF for evaluation; this merge and export precision is separate from the training precision.

5 Experimental Setup

Systems and runtime. Six local systems (base and fine-tuned Qwen3.5 at 0.8B, 2B, 4B) ran as GGUF models via llama-cpp-python with an 8,192-token context on Kaggle dual-T4 instances. Base models ran BF16 GGUF conversions of the official checkpoints, and fine-tuned models ran merged F16 exports, the same at all three scales. GPT OSS 120B (openai/gpt-oss-120b) (OpenAI, 2025) was queried once per condition through the Groq API at low reasoning effort as a single-run informal ceiling, not a replicated baseline.

Conditions. The three conditions differ only in how the “Retrieved Sections” block is filled: empty (*none*); top-5 sections by BM25 (Robertson and Zaragoza, 2009) over the deduplicated section-level corpus of Section 3.1; or top-5 by exact inner-product search (FAISS IndexFlatIP) (Johnson et al., 2017) over paraphrase-multilingual-MiniLM-L12-v2 embeddings (Reimers and Gurevych, 2020). Retrieved context was capped at 1,500 tokens. Both retrievers are deterministic, so across repeated runs the only stochastic component is generation.

Decoding and repetition. All models used temperature 0.7, top-p 0.8, top-k 20, repetition penalty 1.05, and 512 max output tokens. Every condition ran three times with seeds 42, 456, and 2024, applied identically across variants, modes, and exam files.

Prompt design and answer contract. The fine-tuned models reused their training template, whose trained answer convention states the option letter together with the option text (“Therefore, (C) Property only is correct.”). The pairing matters because legal prose is full of parenthesized labels. “Section 12(A)” contains a perfectly good

Model	2022						2023					
	English			Bangla			English			Bangla		
	None	BM25	FAISS	None	BM25	FAISS	None	BM25	FAISS	None	BM25	FAISS
GPT OSS 120B [†]	59	67	72	36	43	42	42	58	57	40	43	41
Qwen3.5 0.8B	7	4	2	0	3	1	3	0	0	0	0	2
Qwen3.5 0.8B FT	22	25	34	2	10	8	10	25	21	8	19	7
Qwen3.5 2B	18	16	18	4	4	6	13	16	17	3	2	3
Qwen3.5 2B FT	18	32	26	13	13	12	11	24	28	9	19	9
Qwen3.5 4B	32	44	46	15	16	14	29	40	38	14	23	12
Qwen3.5 4B FT	23	36	45	16	21	10	19	37	34	14	24	19

Table 2: Strict consistency, the primary score: a question counts as correct only when all three fixed-seed runs answered it correctly, out of 100 questions per cell. Within each base/fine-tuned pair the better value per column is bold (ties both bold). [†]GPT OSS 120B was run once as an informal reference, so repeat scoring does not apply and its row repeats the single-run accuracies.

option letter, and a parser that accepts bare letters will sometimes score the citation, not the answer; only a letter sitting next to its option text is safe to trust. In piloting, base models given the training template did not pick up this convention, so they received an instruction-following prompt that demands the same contract in words: be concise, avoid <think> tags, analyse briefly, and end with exactly one line of the form “Answer: (X) full option text”, with nothing after it. Questions, retrieved sections, decoding parameters, and seeds stayed identical within each base and fine-tuned pair.

Parsing and metrics. A rule-based parser extracts the chosen option letter; unparseable outputs score incorrect, and we keep them with null predictions. We had made the stored predictions with a permissive extractor, so we later went back and checked every one of the 21,600 stored outputs against the letter-plus-text contract. That check confirms 83.5% of the predictions through an adjacent letter-text pair, finds 5.4% unparseable either way, and turns up conflicting letters in 2.4%. We adjudicated a sample of those conflicts by hand and found no systematic winner between the two readings; re-parsing everything the other way changes no qualitative finding (Appendix D). Three scores are reported separately per model, year, language, and condition over the same three runs: **strict consistency** (all three runs correct), **majority vote** (at least two correct), and **mean accuracy** (mean $\pm$ sample SD, $n=3$), each out of 100 questions. We recomputed every table from the per-run, per-question result files, and they matched our stored score report in every cell. For

the primary strict-consistency outcome, we compared the paired binary result for each question within every scale, exam-language, and retrieval cell using a two-sided exact McNemar test. Holm correction controlled family-wise error across all 36 tests.

We also tested the overall effect at each scale using every stored output. A two-sided exact paired randomization test swapped the base and fine-tuned labels once per original exam question, keeping its English and Bangla forms, three retrieval conditions, and three seeded runs in the same cluster. This gave 200 question clusters per scale, containing 1,200 strict question-condition outcomes and 3,600 individual answers for each model variant. Holm correction was applied separately across the three scales for the strict and answer-level families. GPT OSS 120B was excluded because it has one run and no matched fine-tuned variant. Majority-vote comparisons remain descriptive.

6 Results

Table 2 reports the primary strict-consistency grid; Table 3 aggregates it across the twelve matched year, language, and retrieval cells for each model variant. The majority-vote and mean-accuracy tables appear in Appendix B, where Figure 3 also plots all 36 strict-score deltas.

Small models benefit most from fine-tuning. The 0.8B model gains in all twelve strict cells. Its cross-cell mean rises from 1.8 to 15.9, and the single largest jump, from 2 to 34 correct answers, comes on 2022 English with dense retrieval. The 2B story is nearly as one-sided: ten cells improve,

Model	Mean strict accuracy			Retrieval FT vs. base	
	Overall	English	Bangla	gain	FT W-T-L
0.8B base	1.8	2.7	1.0	-1.0	–
0.8B FT	15.9	22.8	9.0	+8.1	12/0/0
2B base	10.0	16.3	3.7	+0.8	–
2B FT	17.8	23.2	12.5	+7.6	10/1/1
4B base	26.9	38.2	15.7	+6.6	–
4B FT	24.8	32.3	17.3	+10.2	4/1/7

Table 3: Aggregates of the strict grid (Table 2); FT = fine-tuned. Mean strict accuracy averages the twelve year-language-retrieval cells (Overall), the six English cells, and the six Bangla cells. Retrieval gain is a model’s average BM25 and FAISS cells minus its no-retrieval cells. FT W-T-L counts the fine-tuned model’s wins, ties, and losses against its base over the twelve cells; the better value in each pair is bold.

one ties, one regresses, and the mean climbs from 10.0 to 17.8. Majority vote and mean accuracy move the same way (Appendix B), so the gains are not just a stability effect.

Fine-tuned models make more of the same retrieval. Retrieval barely moves the base models at these scales. Averaged over BM25 and FAISS, it shifts the 0.8B base model by -1.0 strict points relative to no retrieval; the same retrieved sections shift its fine-tuned counterpart by +8.1. At 2B the split is +0.8 against +7.6. Each pair reads identical context, so the gap cannot come from retrieval quality. The fine-tuned systems do more with the same supplied text.

Fine-tuning does not lift overall 4B accuracy. At 4B the pattern breaks. Fine-tuning lowers the cross-cell mean by 2.1 points and loses seven of twelve strict comparisons. The damage sits in English, where the mean falls from 38.2 to 32.3; Bangla still rises, from 15.7 to 17.3. This base model already turns retrieval into a +6.6 gain, so there was less room to improve, and the 4B comparison shows no consistent benefit from fine-tuning.

Every system scores higher in English. The gap holds in every matched condition, at every scale. Even the single-run GPT OSS 120B reference reaches 72 on 2022 English with dense retrieval but only 42 on the corresponding Bangla exam; that row is an informal ceiling, not a replicated baseline.

Fine-tuning nearly eliminates language drift. Language adherence changes more uniformly than

accuracy. On Bangla exams, the base models answer mostly in English in 44.0% to 53.2% of visible outputs. After fine-tuning, every rate falls below 1% (Table 4). All three reductions survive question-clustered paired testing after Holm correction. Appendix D.1 gives the exact counts and full protocol.

Scale	Base (%)	FT (%)	Δ points	p_{adj}
0.8B	53.2	0.7	-52.5	< .001
2B	44.0	0.2	-43.7	< .001
4B	44.7	0.7	-44.0	< .001

Table 4: Language drift on Bangla exams. Each value is the share of visible outputs written mostly in English. Negative changes favor the fine-tuned model. Tests preserve the nine outputs attached to each original question as one cluster.

Paired tests support gains at 0.8B and 2B, but not 4B. The gains at the two smaller scales survive paired testing; the 4B change does not. The scale-level test keeps translations, retrieval modes, and repeated generations for each of 200 underlying questions in the same cluster. Table 5 shows significant gains on both outcomes at 0.8B and 2B, but neither outcome reaches significance at 4B. Appendix C reports the complementary 36 cell-level McNemar tests.

Scale	Strict consistency		All answers	
	$\Delta/1,200$	p_{adj}	Δ points	p_{adj}
0.8B	+169 (+14.1)	< .001	+7.25	< .001
2B	+94 (+7.8)	< .001	+6.28	< .001
4B	-25 (-2.1)	.266	-0.78	.596

Table 5: Scale-level paired randomization tests clustered by the 200 original exam questions per scale (100 from 2022 and 100 from 2023). “All answers” uses all 3,600 binary outputs per variant and scale. Holm adjustment is across the three scales separately for each outcome family.

7 Discussion

Context-injected fine-tuning helps most where the base model starts weakest. At 0.8B and 2B, base models gain little from retrieval; their fine-tuned counterparts gain 7.6 to 8.1 strict points, and the paired tests of Table 5 confirm the difference is not noise. If retrieval quality alone set the ceiling, both members of a pair should have benefited similarly from identical retrieved text. They did not. At 4B the base model already turns retrieval into a 6.6-point gain, fine-tuning adds no overall accuracy benefit, and the change is not statistically distinguishable from zero. Context injection appears

most useful when the base model has not already picked up that behavior on its own.

These experiments measure output behavior, not mechanism. A correct option does not prove that a model used the supplied statute, and strict consistency mixes correctness with stability. Majority vote and mean accuracy address the stability part; reliance on the statute stays unverified either way.

Several factors could explain the 4B English regressions. The leading suspect is the training recipe: LoRA settings tuned at 0.8B were reused at 4B, and warmup reached 11% rather than the planned 6%. Thinking-mode leakage is the measurable alternative. Across 36 cells, higher leakage accompanies smaller gains (Spearman $\rho = -0.47$; Table 11), and answers trapped inside think blocks usually score as wrong. Short direct-answer supervision may also displace useful behavior where the base model is strongest, which fits the English decline and Bangla improvement. These runs cannot isolate the causes.

The language result is less ambiguous than the accuracy pattern. Fine-tuning drives majority-English replies to Bangla questions below 1% at all three scales (Table 4). This measures response-language adherence, not legal correctness or fine-grained code-switching. The output audit also finds fewer parse failures but new thinking-mode leakage at 2B and 4B (Appendix D). In two traceable cases, a correct option rests on faulty reasoning, while a wrong option includes fluent but fabricated law (Appendix G). None of these diagnostics estimates legal reliability.

8 Conclusion

Context-injected fine-tuning produced clear gains at 0.8B and 2B on the two bilingual Bangladesh Bar Council exams, and both survive paired significance testing. At those scales the fine-tuned models also did much more with the same retrieved statutory context than their base counterparts. Across all three scales, it reduced majority-English replies to Bangla questions from rates between 44.0% and 53.2% to rates below 1%. The effect on answer language is consistent even though the 4B accuracy result is not. At 4B, Bangla accuracy improves modestly while several English conditions regress, leaving no detectable net effect. The released corpus, 2,165-record dataset, and per-question outputs make the

comparison auditable and reproducible.

Acknowledgments

This work was carried out as a Directed Research project at the Department of Electrical and Computer Engineering, North South University.

Limitations

The GGUF formats differ slightly, and the 4B base evaluation ran in a Kaggle notebook environment that was not itself preserved, though its notebook source and per-question outputs were. The LoRA and optimization configuration was selected for the 0.8B model and reused unchanged at 2B and 4B; we suspect the larger models needed scale-specific settings, and that reusing this recipe contributes to the 4B English regressions. The effective batch (32) and warmup share (11%) we ran also differ from the targets we configured (16 and 6%), as reported in Section 4. The permissive answer parser can confuse option letters in citations or restated options with the final choice; our letter-plus-text audit of all stored outputs bounds such conflicts at 2.4%, with no systematic direction under adjudication (Appendix D).

Multiple-choice accuracy is a narrow proxy for legal QA, while strict consistency also rewards cross-seed stability. Cell-level McNemar tests and scale-level question-clustered randomization tests assess the primary strict outcome; the latter also tests all individual answers without assuming that translations, retrieval modes, or seeded repetitions are independent. These tests were added after inspection of the results and were not preregistered. The two exam years and 200 underlying questions still limit power and generalization; majority-vote comparisons remain descriptive. The English sets are machine translations without independent legal review. The 2022 and 2023 benchmark questions were withheld from synthetic-data generation, and the exact-match audit found no evaluation question copied into training. Possible exposure of the pretrained base checkpoints to these public examinations is outside our data pipeline and was not auditable. Repeated development use of the same 400 questions also creates benchmark co-adaptation risk.

The study covers six acts, three schedules, two exam years, one model family, and MCQs only. It does not establish performance across Bangladeshi law or on open-ended legal work.

The corpus and QA data were partly LLM-generated; the authors reviewed them, but no per-item review record supports a claim of exhaustive legal verification. Preserved and missing artifacts are listed in Appendix E.

Ethics Statement

These systems are research artifacts, not legal-advice tools. Even the best local model answers most Bangla questions incorrectly under strict scoring, and inspected outputs include confident fabricated citations. Deployment would require qualified human oversight, provenance checks, and broader evaluation. The released datasets contain public statutory and exam text and no personal data.

References

- Sabik Aftahee, A. F. M. Farhad, Arpita Mallik, Ratnajit Dhar, Jawadul Karim, Nahiyen 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.
- Arif Ahmed. 2021. [Challenges with access to justice in Bangladesh](#). Centre for Governance Studies. Commentary. Accessed 2026-06-23.
- Farzana Akter. 2017. [Legal aid for ensuring access to justice in Bangladesh: A paradox?](#) *Asian Journal of Law and Society*, 4(1):257–275.
- Shimanto Bhowmik, Tawsif Tashwar Dipto, Md Sazzad Islam, Sheryl Hsu, and Tahsin Reasat. 2025. [Evaluating LLMs’ multilingual capabilities for Bengali: Benchmark creation and performance analysis](#). *Preprint*, arXiv:2507.23248.
- Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. [SaulLM-7B: A pioneering large language model for law](#). *Preprint*, arXiv:2403.03883.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). *Preprint*, arXiv:2305.14314.
- Government of the People’s Republic of Bangladesh. 1972. The Constitution of the People’s Republic of Bangladesh, Article 27. <http://bdlaws.minlaw.gov.bd/>. Laws of Bangladesh portal, Ministry of Law, Justice and Parliamentary Affairs. Accessed 2026-07-16.
- Daniel Han and Michael Han. 2024. [Unsloth: Efficient LLM fine-tuning](#). Software repository, version unspecified. Accessed 2026-07-16.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [LoRA: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Md Jahirul Islam, Masahiro Suzuki, and Nurunnahar Mazumder. 2024. [Promoting access to justice in Bangladesh: Towards a hybrid justice model](#). *International Journal of Law, Crime and Justice*, 77:100655.
- Arafat Reza Jaan. 2023. [What makes access to Justice elusive in Bangladesh?](#) LSE South Asia Blog. Blog analysis; cited URL is the BRAC University CPJ reprint. Accessed 2026-06-22.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. [Billion-scale similarity search with GPUs](#). *Preprint*, arXiv:1702.08734.
- Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. [IL-TUR: Benchmark for Indian legal text understanding and reasoning](#). *Preprint*, arXiv:2407.05399.
- Muhammad Rafsan Kabir, Rafeed Mohammad Sultan, Fuad Rahman, Mohammad Ruhul Amin, Sifat Momen, Nabeel Mohammed, and Shafin Rahman. 2025. [LegalRAG: A hybrid RAG system for multilingual legal information retrieval](#). *Preprint*, arXiv:2504.16121.
- Jihyung Lee, Daehui Kim, Seonjeong Hwang, Hyounghun Kim, and Gary Lee. 2025. [KoBLEX: Open legal question answering with multi-hop reasoning](#). *Preprint*, arXiv:2509.01324.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems* 33.
- Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. 2024. [Keeping LLMs aligned after fine-tuning: The crucial role of prompt templates](#). *Preprint*, arXiv:2402.18540.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- 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.

- Nils Reimers and Iryna Gurevych. 2020. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). *Preprint*, arXiv:2004.09813.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: BM25 and beyond](#). *Foundations and Trends in Information Retrieval*, 4(1–2):1–174.
- Umme Sharaban Tahura and Shawkat Alam. 2025. [Legal aid, systemic barriers, and access to justice in Bangladesh](#). *The Daily Star*. News analysis. Accessed 2026-06-23.
- K. Veningston and Apratim Mishra. 2025. [Dataset for legal question answering system in the Indian judiciary context](#). *Data in Brief*, 60:111647.
- Azmine Touseh Wasi, Wahid Faisal, Mst Rafia Islam, and Mahathir Mohammad Bappy. 2024. [Exploring possibilities of AI-powered legal assistance in Bangladesh through large language modeling](#). *Preprint*, arXiv:2410.17210.
- Azmine Touseh Wasi, Wahid Faisal, Mst Rafia Islam, and Md Rizwan Parvez. 2025. [Mina: A multilingual LLM-powered legal assistant agent for Bangladesh for empowering access to justice](#). *Preprint*, arXiv:2511.08605.
- Sabina Yesmin. 2025. [Alternative dispute resolution: A strategic solution to Bangladesh’s judicial backlog](#). *The Business Standard*. News analysis. Accessed 2026-06-23.

A Dataset and Training Details

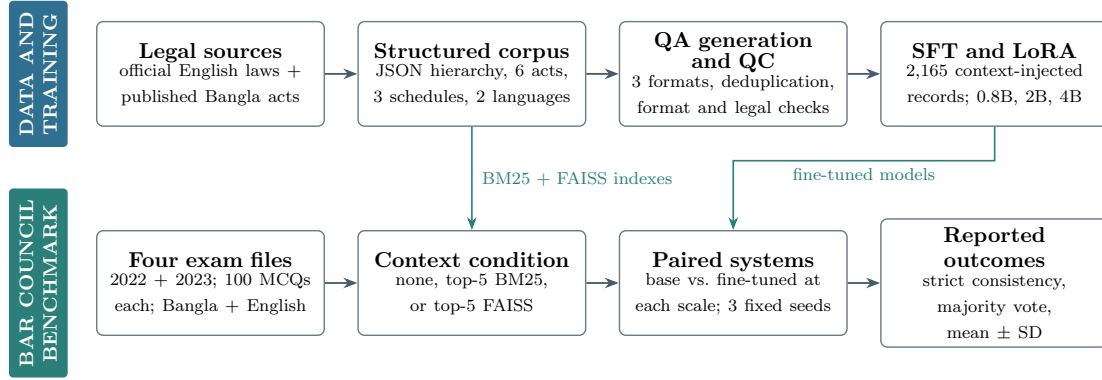


Figure 2: Detailed artifact flow. The upper row builds context-injected fine-tuning data and models; the lower row evaluates base and fine-tuned systems on four Bangladesh Bar Council exam files. The structured corpus also supplies the BM25 and FAISS retrieval indexes.

Component	Setting
Training method	LoRA via Unsloth; 4-bit loading requested, not confirmed by logs
LoRA	$r = 32$, $\alpha = 64$, dropout 0.05; seven linear projections
Optimizer	AdamW 8-bit, learning rate 2×10^{-4} , cosine decay
Schedule	2 epochs, weight decay 0.01, cosine warmup (14 steps)
Batch and sequence	effective batch 32 (executed; config targeted 16); 2,048 tokens
Split	1,948 train / 217 validation, seed 42
Inference	temperature 0.7, top-p 0.8, top-k 20
Generation	repetition penalty 1.05; 512 output tokens
Runs	seeds 42, 456, 2024
Retrieval	top-5 BM25 or FAISS; 1,500-token context cap

Table 6: Training and inference configuration.

B Full Results

Table 2 in the main text reports the primary strict-consistency grid. Tables 7 and 8 complete it with majority-vote and mean-accuracy scoring over the same runs.

Model	2022						2023					
	English			Bangla			English			Bangla		
	None	BM25	FAISS	None	BM25	FAISS	None	BM25	FAISS	None	BM25	FAISS
GPT OSS 120B [†]	59	67	72	36	43	42	42	58	57	40	43	41
Qwen3.5 0.8B	35	30	32	22	18	17	23	24	26	12	22	14
Qwen3.5 0.8B FT	40	45	53	18	24	25	27	39	39	27	33	23
Qwen3.5 2B	39	47	42	20	20	11	28	36	34	14	18	13
Qwen3.5 2B FT	43	50	52	28	31	32	28	37	38	25	27	25
Qwen3.5 4B	47	58	59	27	36	28	47	53	50	38	37	32
Qwen3.5 4B FT	45	59	63	35	37	30	42	52	53	36	43	38

Table 7: Majority vote: at least two of three runs are correct. Each cell contains 100 questions. [†]Single run.

Model	2022						2023					
	English			Bangla			English			Bangla		
	None	BM25	FAISS	None	BM25	FAISS	None	BM25	FAISS	None	BM25	FAISS
0.8B	36.0±8.5	37.0±9.5	34.7±11.0	23.7±9.5	23.0±5.0	22.0±10.6	28.7±1.2	29.7±13.8	30.0±12.1	21.3±2.3	22.3±10.7	22.7±6.4
0.8B FT	42.3±3.1	43.7±2.3	52.7±2.3	23.0±5.0	25.7±1.5	29.0±6.2	30.3±1.5	41.3±3.5	38.7±1.2	30.0±6.1	33.3±1.2	28.0±3.0
2B	41.3±0.6	43.0±5.3	43.3±4.5	25.7±2.1	27.0±4.4	21.0±7.0	33.3±2.3	38.3±3.1	35.7±4.0	19.3±3.5	24.0±3.5	18.3±4.7
2B FT	45.7±4.0	50.7±1.5	47.7±4.7	33.3±3.8	34.0±5.2	31.7±2.3	34.0±4.6	39.7±2.1	40.7±4.5	26.7±3.8	32.0±0.0	29.7±6.1
4B	51.7±3.5	58.7±3.2	61.3±4.0	35.7±6.7	40.3±1.5	33.0±2.6	50.0±2.6	56.0±2.6	53.7±3.2	39.0±4.0	41.7±4.5	34.3±1.5
4B FT	47.0±4.6	57.7±1.5	62.7±0.6	37.7±0.6	41.0±7.0	33.0±2.0	41.3±1.2	55.0±5.6	54.3±3.5	34.7±4.6	43.0±2.6	38.7±2.9

Table 8: Mean per-run accuracy $\pm$ sample standard deviation ($n = 3$) for the six Qwen3.5 systems (model size, base or fine-tuned).

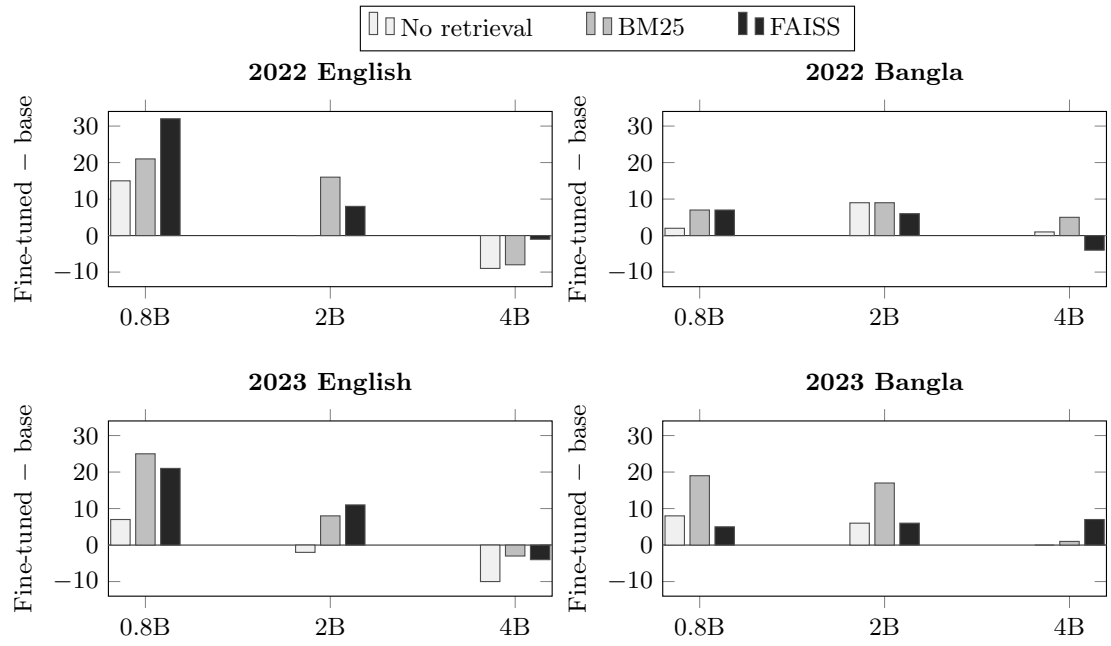


Figure 3: Fine-tuned minus base strict-consistency score (out of 100) for every scale, year, language, and retrieval condition, computed from Table 2. No retrieval leaves the context block empty; BM25 and FAISS fill it with the top-5 retrieved sections. All panels share the same y-axis scale; bars below zero are regressions.

C Cell-Level Paired Significance Tests

Figure 4 gives the secondary cell-level analysis behind the scale-level tests in Table 5. Each McNemar test uses the same 100 questions for the base and fine-tuned variants. Holm adjustment controls family-wise error across all 36 cells.

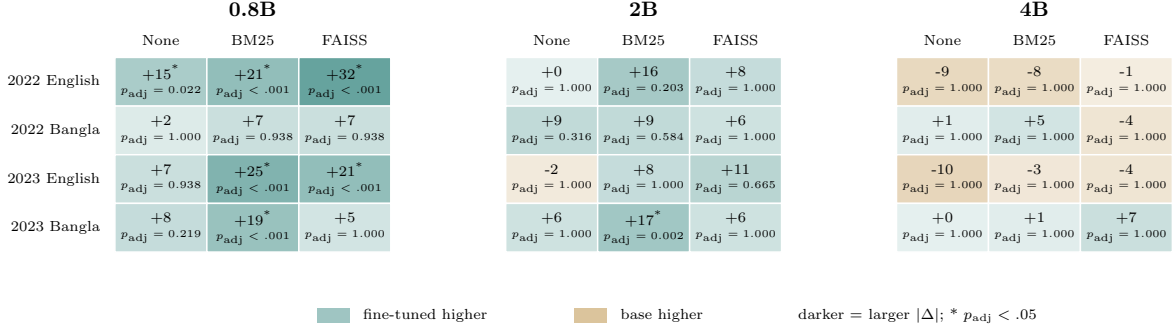


Figure 4: Fine-tuned-minus-base strict-consistency difference and Holm-adjusted p value for each scale, exam language, year, and retrieval condition. Stars mark adjusted $p < .05$.

D Error Prevalence and Parser Audit

Table 9 counts three failure modes over every stored output, 1,800 per row: two exam years, three retrieval conditions, and three runs of 100 questions. Parse failure means the stored prediction is null. Think leakage means the output contains a model-emitted `<think>` block. Language drift is computed on Bangla exams over outputs with visible non-think text: the share whose visible answer is majority non-Bangla script.

Scale	Var.	Parse fail %		Think leak %		Drift %
		Bn	En	Bn	En	
0.8B	base	19.1	8.4	0.0	0.0	53.2
0.8B	FT	0.8	0.2	2.9	3.7	0.7
2B	base	26.1	16.7	0.0	0.0	44.0
2B	FT	0.2	0.8	0.1	30.7	0.2
4B	base	7.6	5.0	0.0	0.0	44.7
4B	FT	2.3	0.7	33.1	22.8	0.8

Table 9: Prevalence of countable failure modes per scale, variant, and exam language, aggregated over years, retrieval conditions, and runs (analysis/error_prevalence.py).

D.1 Paired Analysis of Language Drift

Table 4 gives the main rates and effect sizes. Table 10 adds their exact counts and analysis details. One cluster contains the nine Bangla outputs for an original exam question: three retrieval conditions by three seeds. Each of 200,000 paired randomizations swaps the base and fine-tuned labels for whole question clusters. The statistic is the difference in majority-English rates among outputs with visible non-think text, recomputed after every swap. Holm adjustment is across the three scales.

Scale	Base drift	FT drift	Δ points	p_{adj}
0.8B	819/1,540 (53.2%)	12/1,747 (0.7%)	-52.5	< .001
2B	764/1,738 (44.0%)	4/1,795 (0.2%)	-43.7	< .001
4B	792/1,771 (44.7%)	9/1,204 (0.7%)	-44.0	< .001

Table 10: Language drift on Bangla exams. Counts are majority-English visible outputs divided by outputs with visible non-think text. Tests are two-sided, question-clustered paired randomization tests (analysis/language_drift_significance.py).

Table 9 aggregates over retrieval conditions and years; the leakage-delta correlation reported in Section 7 (Spearman $\rho = -0.47$) instead uses the per-cell fine-tuned leakage rates of Table 11, one rate per scale, exam, and retrieval condition, paired with that cell’s fine-tuned-minus-base strict delta (analysis/leakage_correlation.py).

Scale	Condition	2022		2023	
		English	Bangla	English	Bangla
0.8B	None	1.7	5.0	4.0	4.3
0.8B	BM25	5.7	0.7	4.7	1.3
0.8B	FAISS	2.3	4.0	4.0	2.3
2B	None	34.3	0.3	38.3	0.3
2B	BM25	26.7	0.0	43.0	0.0
2B	FAISS	21.7	0.0	20.0	0.0
4B	None	39.3	16.0	39.7	13.0
4B	BM25	10.7	38.0	18.0	36.7
4B	FAISS	11.3	53.0	18.0	42.0

Table 11: Fine-tuned thinking-mode leakage rate (%) per evaluation cell: share of the 300 outputs (three runs of 100 questions) containing a model-emitted <think> block.

The stored predictions come from the original permissive extractor, so we went back and re-checked all 21,600 of them against the letter-plus-text contract of Section 5 (analysis/pair_validate.py). The pairing confirms 83.5%. Another 5.4% parse under neither reading, 6.8% carry a stored letter the pairing cannot validate, and 2.0% validate where the stored prediction is null; the remaining 2.4% disagree outright. We read a stratified sample of forty disagreements by hand. Twelve favored the stored reading, ten the pairing, seventeen were ambiguous or cut off before any verdict, and one matched neither; exactly one of the forty came from a misread citation. The reported tables keep the stored predictions. Re-parsing everything the other way moves the affected cells by a few points, in directions that leave every qualitative finding intact and, if anything, slightly enlarge the 4B English regressions.

E Reproducibility Inventory

The complete public release, including all code, data, corpus, fine-tuning files, and per-question evaluation outputs, is available at <https://github.com/m-mahadi/bangladesh-legal-qa>. The project archive contains the structured law corpus, six QA splits totaling 2,165 records, two chat-format SFT files, and four evaluation files. Executed training notebooks and trainer summary logs are preserved for all three scales, 0.8B, 2B, and 4B; step-level metric histories are not. Executed evaluation notebook sources and per-question outputs are preserved locally for all three scales, 0.8B, 2B, and 4B. We recomputed the three reported tables from those outputs, and they matched our stored score report. The paired significance analysis is preserved as analysis/strict_mcnemar.py; its 36-cell output is analysis/strict_mcnemar_holm.csv, and its scale-level output is analysis/scale_level_significance.csv. The parser audit, error-prevalence analysis, leakage-delta correlation, and language-drift test are preserved as analysis/pair_validate.py, analysis/error_prevalence.py, analysis/leakage_correlation.py, and analysis/language_drift_significance.py with their CSV outputs. Missing artifacts include the executed GPT OSS evaluation notebook and per-item human-review records.

F Original Bangla Prompt Artifacts

The final Bangla generation prompts survive as retained UTF-8 source files from the original report workflow and are rendered directly from those files, preserving the authors’ wording and Bengali script. Each prompt spans two pages so the text can be inspected at normal page size.

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. The model must reason through each answer using a three-step reasoning chain (Issue, Rule, Application). All IRAC_Reasoning values must be written in Bangla. Answer is the direct output after Application and is the only user-facing field.

[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 — not a demonstration of your own capability. Smaller models learn by pattern recognition across many examples: they need structural consistency, not sophistication. A 100-word IRAC block gives a small model nowhere to anchor the reasoning pattern; a tight 75-word block with predictable structure and clean logical flow does. Length variance across entries forces the model to learn noise instead of structure. Format drift between entries — different openers, inconsistent sentence counts, uneven field lengths — degrades SFT quality directly. Every constraint in this prompt exists to produce data the model can generalize from, not data that reads impressively in isolation.

[Jurisdiction & Corpus]
Jurisdiction : [Bangladesh / India / Pakistan / UK] — select one.
Legal Corpus : Use only the Act names, section numbers, and legislative addresses that appear in the provided dataset. Do not import terminology or citation formats from other jurisdictions.
Citation Note : Legislative addresses vary by instrument. Use whichever structure the provision actually uses — a Section, an Order and Rule, a Schedule and Article, or a combination. Never invent a Section number if the provision is an Order, Rule, or Article.

[Scope of Work]
Target Range: Generate questions for Sections [Start Section] to [End Section].
Volume: Provide [Number] question(s) per section / Provide [Number] total questions.

[Output Style — SFT Quality Targets]
These targets apply to every entry. They exist to produce concise, consistent, high-signal SFT data.
- IRAC_Reasoning total length: 45–75 words (in Bangla) across all three fields combined.
- Answer field: 25–45 words maximum.
- Do not pad. Do not elaborate beyond what is legally necessary. Every word must earn its place.

[Format & Schema]
Output must be a strictly valid JSON array of objects.
Schema per object:
Section Number : The specific legislative address of the provision (e.g., "Section 1", "Order XXXIX Rule 1", "Schedule I Article 23").
Entry_ID : চূড়ান্ত generation run-এর মধ্যে entry-র একটি ক্রমিক সূচিপত্র যা পন্যক্তকারী (যেমন, 1, 2, 3)। একটি নির্দিষ্ট key নাম ব্যবহার করুন — সংঘটিত key-এর মধ্যে এখেন্ড করবেন না।
Question : একটি স্পষ্ট, যথার্থিত ভাষায় উদ্ভূত প্রশ্নের জন্য "Scenario-based" ব্যবহার করুন। একটি সাধারণ মানুষের ভাষায় (client-speak) লিখতে হবে — একজন মর্মান ভেতবে তার সমস্যা বর্ণনা করেন সেভাবে। আইনি পরিভাষা IRAC_Reasoning এবং Answer-এ ব্যবহার করতে হবে, প্রস্তুত নয়।
Subsection/Clause : The specific sub-provision or clause the question focuses on (e.g., "Subsection (2)", "Clause (c)", "Rule 1(a)"). Use "N/A" or "General Provision" for single-paragraph provisions.
Section Text : The full, verbatim text of the section from the dataset (the whole section, not only the clause or subsection).
IRAC_Reasoning : সকল মান বালায় তিনি key সহ একটি object — Issue, Rule, Application। Conclusion field নেই — Answer হলো Application-এর পরের সরাসরি output। শুধুমাত্র অভ্যন্তরীণ মান-নিয়ন্ত্রণের জন্য। কোনো chatbot বা UI-তে ব্যবহারকারীকে এটি দেখানো যাবে না।
Answer : ব্যবহারকারীর জন্য একমাত্র দৃশ্যমান উত্তর। নিচের Answer Field নিয়মাবলী দেখুন।
NO_IRAC_Reasoning : IRAC কাঠামো ছাড়া সহজ ব্যাখ্যা। নিচের NO_IRAC_Reasoning Field নিয়মাবলী দেখুন।
Type : নির্দেশ Type Selection Method অনুসরণ করে একটি type নির্বাচন করুন।
Difficulty : Assign a level (Easy, Medium, or Hard).
Keywords : 3–5 legal terms or concepts central to the section, in Bangla.
Cited Acts and Sections : The primary Act name (in English) and the specific legislative address. Must be the last key.

[IRAC_Reasoning Keys — সকল মান বাংলায় লিখতে হবে]
Note: IRAC_Reasoning হলো Answer-এ পৌছানোর অভ্যন্তরীণ reasoning chain। তিনটি field: Issue, Rule, Application। Conclusion field নেই — Answer সেই ভূমিকা পালন করে। মান-নিয়ন্ত্রণের জন্য JSON-এ সংরক্ষিত থাকবে, কিন্তু সকল UI/chatbot-এ ব্লকিয়ে রাখতে হবে।

"Issue" : এক বাক্যে। পরিস্থিতিতে উপস্থাপিত সূনির্দিষ্ট আইনি প্রশ্নটি বুলন। ঘটনার বিবরণ বাদ দিয়ে শুধু মূল আইনি সমস্যাদি চিহ্নিত করুন।
"Rule" : সর্বোচ্চ এক বাক্য। সংশ্লিষ্ট ধারার মূল আইনি নীতিটি নিজেই ভাষায় প্যারাজেজ করুন। Section Text থেকে কোনো ব্যাক্যাংশ ছব্ব ব্যবহার করবেন না। আইনটিতে তার সবচেয়ে গুরুত্বপূর্ণ কার্যকরী অর্থে সংক্ষিপ্ত করুন।
"Application" : সর্বোচ্চ এক থেকে দুই বাক্য। Rule-টি পরিস্থিতির সূনির্দিষ্ট তথ্যের সাথে প্রয়োগ করুন। ধারার কোন উপাদানটি সক্রিয় হয়েছে এবং কেন উত্তরাটি সেখান থেকে আসে তা বুলন। অপ্রয়োজনীয় বিস্তার এড়িয়ে চলুন। প্রতিটি entry-তে Application বাক্যের শুরু পঠন পরিবর্তন করুন — একই phrase বা বাক্য-গঠন দিয়ে প্রতিটি entry শুরু করবেন না।

বিশেষ দৃষ্টব্য: Application-এর পরে reasoning chain সম্পূর্ণ। Answer হলো plain-language output — এটি কোনো চতুর্থ IRAC field নয়। JSON-এ Conclusion key যোগ করবেন না।

[Answer Field — ব্যবহারকারীর জন্য একমাত্র দৃশ্যমান উত্তর]
Answer ক্ষেত্রটি IRAC_Reasoning-এর ঠিক পরে স্থাপন করতে হবে।

নিয়মাবলী:
- Act-এর নাম (ইংরেজিতে) দিয়ে শুরু করুন, তারপর সঠিক legislative address। যেমন:
Section : "The Evidence Act, 1872, Section 22, Clause (1) অনুযায়ী, ..."
Order and Rule : "The Code of Civil Procedure, 1908, Order XXXIX, Rule 1 অনুযায়ী, ..."
Schedule/Article : "The Limitation Act, 1908, Schedule I, Article 23 অনুযায়ী, ..."
Combined : Act-এ যে কোনো আছে সেই ক্রমে সকল প্রয়োজ্য ভগ (Subsection / Clause / Illustration / Rule) অন্তর্ভুক্ত করুন।
- কখনো শুধু section number বা provision নাম দিয়ে শুরু করবেন না।
- সহজ, সরল বাংলায় — প্রশ্নটি সেভাবে ভিত্তিকেস করা হয়েছে সেভাবে উত্তর দিন।
- সর্বোচ্চ এক থেকে দুই বাক্য। ২৫–৪৫ শব্দ।
- উদ্ভূতির জন্য প্রয়োজনীয় পরিভাষার ব্যতীে আইনি জটিল ভাষা পরিহার করুন।

উদাহরণ:
"Answer": "The Code of Criminal Procedure, 1898, Section 4A(2)(b) অনুযায়ী, লাইসেন্স প্রদান বা বাতিলের মতো প্রশাসনিক কাজ — তাই এই সিদ্ধান্ত নির্বাহী ম্যাজিস্ট্রেট নেবেন।"

[NO_IRAC_Reasoning Field — ব্যবহারকারীর জন্য নয়, training signal হিসেবে]
Answer-এর ঠিক পরে স্থাপন করতে হবে।

নিয়মাবলী:
- দুই থেকে তিন বাক্য সহজ বাংলা — Issue / Rule / Application / Conclusion সেভাবে ব্যবহার করা যাবে না।
- সম্পূর্ণ legislative citation অথবা prose-এর মধ্যে ব্যাবহিকভাবে অন্তর্ভুক্ত করতে হবে — Act-এর নাম এবং legislative address। Citation বাদ দেওয়া যাবে না। IRAC_Reasoning এবং NO_IRAC_Reasoning-এর মধ্যে, একমাত্র পার্থক্য হলো কাঠামোবদ্ধ লেবেলের উপস্থিতি বা অনুপস্থিতি, citation-এর নয়।
- একজন জ্ঞানী ব্যক্তি যেভাবে একজন সাধারণ মানুষকে ব্যাখ্যা করবেন সেভাবে লিখুন।
- IRAC_Reasoning-এর মতো একই আইনি বিষয় করার করতে হবে, কিন্তু কাঠামোবদ্ধ ফরম্যাট ছাড়া।
- Section Text ছব্ব প্রস্তুতস্থাপন করবেন না। শুধু প্যারাজেজ করুন।
- লক্ষ্যমাত্রা: ৩০–৫০ শব্দ।

উদাহরণ:
"NO_IRAC_Reasoning": "The Code of Criminal Procedure, 1898, Section 4A(2)(b) অনুযায়ী, লাইসেন্স প্রদান বা বাতিলের মতো প্রশাসনিক কাজ বিচারিক ম্যাজিস্ট্রেটের এখতিয়ারে পড়ে না। এই ধারটি স্পষ্টভাবে এই ধরনের প্রশাসনিক সিদ্ধান্তের ক্ষমতা নির্বাহী ম্যাজিস্ট্রেটকে দিয়েছে।"

[Type Selection Method]
সময়ক ফাইল QUESTION_TYPE VOCABULARY.md দেখুন।
প্রক্রিটি প্রশ্ন তৈরির সময় নিম্নলিখিত প্রক্রিয়া অনুসরণ করুন:
Step 1 — QUESTION_TYPE VOCABULARY.md ফাইলের শেষে Quick Reference তালিকা থেকে একটি type এলামেন্টাভাবে নির্বাচন করুন।
Step 2 — যাচাই করুন যে নির্বাচিত type-টি প্রক্রিয়াজনক ধারায় প্রকৃতপক্ষে পরীক্ষা করা যায়। যদি হ্যাঁ, সেই ব্যবহার করুন। যদি না হয়, পরবর্তী সবচেয়ে উপযুক্ত type নির্বাচন করুন।
Step 3 — প্রথমে প্রশ্ন লিখে পরে type নির্বাচন করবেন না — বরং type নির্বাচন করে তারপর সেই type-এর উপযোগী প্রশ্ন ও পরিস্থিতি তৈরি করুন।
Step 3 — যেকোনো ১০টি প্রশ্নের ব্যাচে কমপক্ষে ৬টি ভিন্ন type নিশ্চিত করুন। একই type পুনরাবৃত্তি হবার বেশি ব্যবহার করবেন না।

[Constraint Guidelines]
Verbatim Integrity : "Section Text" ক্ষেত্রটি JSON ভেউসেট থেকে ছব্ব কপি করতে হবে। ব্লক থেকে পুনর্নির্মাণ বা সংশোধন করা যাবে না। যদি ভেউসেট থেকে সঠিক পাঠ পাওয়া না যায়, অনুমান করবেন না — entry-টি incomplete হিসেবে চিহ্নিত করুন।
Anti-Hallucination : যাচাই থেকে করবেন ধারার পাঠ পুনর্নির্মাণ বা প্যারাজেজ করবেন না। সকল "Section Text" অবশ্যই প্রশ্ন ভেউসেট থেকে আসতে হবে।
Structural Order : "Section Number" প্রথমে এবং "Cited Acts and Sections" শেষ key হবে।
Granularity : ধারায় উপধারা থাকলে সর্বদা "Subsection/Clause" পূরণ করতে হবে। একক অনুচ্ছেদের ধারার জন্য "N/A" বা "General Provision" ব্যবহার করুন।
Accuracy : শুধুমাত্র প্রশ্ন পাঠের মধ্যে থাকুন। ভেউসেটে নেই এমন কোনো পাঠ বা ক্রস-রেফারেন্স উদ্ভাবন করবেন না।
IRAC Rule Step : শুধু প্যারাজেজ করুন। IRAC_Reasoning-এর ভেউসেট Section Text থেকে কোনো ব্যাক্যাংশ ছব্ব ব্যবহার করবেন না।
Client-Speak Rule : প্রশ্নটি অবশ্যই একজন সাধারণ মানুষের ভাষায় হতে হবে — প্রদত্ত মধ্যে আইনি পরিভাষা ব্যবহার করবেন না।
Length Discipline : কোনো IRAC field তার শব্দসীমা অতিক্রম করলে আউটপুটের আগে সংশোধন করুন। সংক্ষিপ্ততা, একটি মানের প্রয়োজনীয়তা।
Answer Visibility Rule: IRAC_Reasoning কোনো ব্যবহারকারীমুখী interface বা chatbot-এ প্রদর্শন করা যাবে না। শুধুমাত্র Answer ক্ষেত্রটি ব্যবহারকারীকে দেখানো হবে।
Scenario Freshness : প্রতিটি scenario-তে অন্যান্য factual context ব্যবহার করতে হবে। পূর্বের entry-তে একই চরিত্রের নাম, চরিত্রের বর্ণন বা বিবরণের পরিষ্কৃত পুনরাবৃত্তি ব্যবহার করবেন না।
Multi-Clause Sections : কোনো ধারায় একাধিক clause থাকলে, সবচেয়ে আইনভাবে গুরুত্বপূর্ণ বা ব্যবহারিকভাবে প্রাসঙ্গিক clause-টিকে কেন্দ্র করে একটি প্রশ্ন তৈরি করুন। প্রতিটি clause-এর জন্য অন্যান্য প্রশ্ন তৈরি করবেন না। Subsection/Clause field-এ লক্ষ্যভিত্তিক clause উল্লেখ করুন।

[Pre-Output Quality Check — আউটপুট দেওয়ার আগে প্রতিটি এক্সির জন্য নিচের সকল বিষয় যাচাই করুন।]
□ "Section Text" — ভেউসেট থেকে ছব্ব কপি করা হয়েছে, যাচাই থেকে পুনর্নির্মাণ করা হয়নি
□ "Rule" — Section Text থেকে কোনো ব্যাক্যাংশ সরাসরি কপি করা হয়নি, এক বাক্য
□ "Application" — এক থেকে দুই বাক্য, ধারার নির্দিষ্ট কোন উপাদানটি সক্রিয় হয়েছে তা চিহ্নিত করা হয়েছে
□ "Answer" — Act নাম দিয়ে শুরু হয়েছে, সম্পূর্ণ legislative address আছে; ২৫–৪৫ শব্দ; সহজ বাংলায়
□ "NO_IRAC_Reasoning" — দুই থেকে তিন বাক্য; IRAC লেবেল নেই; Act নাম ও legislative address prose-এ অন্তর্ভুক্ত; ৩০–৫০ শব্দ; শুধু প্যারাজেজ

Bangla single-hop generation prompt, part 1 of 2.

```
□ "Type" — QUESTION_TYPE_VOCABULARY.md থেকে নির্বাচন করা হয়েছে, উদ্ধারিত বা ফাঁকা নয়
□ প্রশ্নের ভাষা — সাধারণত মন্তব্যের ভাষায়, আইনি পরিভাষামূলক
□ যেটি IRAC_Reasoning পদ্ধতিগত — তিনটি field নিম্নে ৪৫-৭০ শব্দ
□ Scenario freshness — এই batch-এর অন্য সব entry থেকে factual context আলাদা; একই চরিত্রের নাম, লেনদেনের ধরন বা ঘিরেধের পরিষ্টিত পুনরায় ব্যবহার করা হয়নি

[Negative Example — What to Avoid]
The following is a FAILING version of a single-hop entry. Annotated errors show what violates the rules. Do not produce entries like this.

"IRAC_Reasoning": {
  "Issue": "ফৌজদারি কার্যবিধি কী করতে পারে?",
  // X FAIL — Issue: অত্যন্ত অস্পষ্ট। সুনির্দিষ্ট আইনি প্রশ্নটি চিহ্নিত করা হয়নি। এটি প্রশ্নের পুনরাবৃত্তি মাত্র।

  "Rule": "ফৌজদারি কার্যবিধি সমগ্র বাংলাদেশে প্রযোজ্য হইবে এবং বিশেষ আইনকে প্রভাবিত করিবে না।",
  // X FAIL — Rule: Section Text থেকে সরাসরি অনূদয়/কপি করা হয়েছে। Rule অবশ্যই নিজের ভাষায় paraphrase হতে হবে — Section Text-এর কোনো বাক্যাংশ ব্যবহার করা যাবে না।

  "Application": "এই ক্ষেত্রে ধারা ১(২) প্রযোজ্য। বিশেষ আইনটি আলাদা পদ্ধতি নির্ধারণ করেছে। ফৌজদারি কার্যবিধিতে কোনো বিপরীত বিধান নেই। তাই বিশেষ পদ্ধতি বহাল থাকবে।",
  // X FAIL — Application: চার বাক্য — সর্বোচ্চ দুই বাক্য। Fragment structure reasoning-এর যোগসূত্র নষ্ট করে। Application অবশ্যই trigger condition থেকে পরিণতিতে একটি অবিচ্ছিন্ন যুক্তিতে পৌঁছাতে হবে।
}

Corrected version: see the Example below.

Example:
{
  "Section Number": "Section 1",
  "Entry_ID": 1,
  "Question": "Scenario-based: একটি বিশেষ আইনে নির্দিষ্ট ধরনের মামলার জন্য আলাদা বিচার পদ্ধতি রয়েছে। ফৌজদারি কার্যবিধি কি সেই পদ্ধতিকে বাতিল করে দিতে পারে?",
  "Subsection/Clause": "Subsection (2)",
  "Section Text": "(১) এই আইন ফৌজদারি কার্যবিধি আইন, ১৮৯৮ নামে অভিহিত হইবে, এবং ইহা ১৮৯৮ সালের ১লা জুলাই তারিখে কার্যকর হইবে। (২) ইহা সমগ্র বাংলাদেশে প্রযোজ্য; কিন্তু, বিপরীত কোন সুনির্দিষ্ট বিধান না থাকিলে, এখানে অন্তর্ভুক্ত কোন কিছুই সর্বোচ্চমতে কার্যকর বিশেষ আইন, বা অন্য কোন আইন দ্বারা প্রদত্ত বিশেষ প্রত্যয়র বা ক্ষমতা, বা নির্ধারিত বিশেষ পদ্ধতিকে প্রভাবিত করিবে না।",
  "IRAC_Reasoning": {
    "Issue": "বিশেষ আইনের বিচার পদ্ধতিকে ফৌজদারি কার্যবিধি প্রভাবিত করতে পারে কিনা?",
    "Rule": "ফৌজদারি কার্যবিধি সারা বাংলাদেশে প্রযোজ্য হলেও বিপরীত সুনির্দিষ্ট বিধান না থাকলে অন্য কোনো বিশেষ আইনের পদ্ধতি বা ক্ষমতাকে এটি প্রভাবিত করে না।",
    "Application": "প্রশ্নটি সরাসরি ধারা ১(২)-এর অওগার পড়ে — বিশেষ আইনের পদ্ধতি সুরক্ষিত, যদি না কার্যবিধিতে বিপরীত বিধান থাকে।",
  },
  "Answer": "The Code of Criminal Procedure, 1898, Section 1(2) অনুযায়ী, ফৌজদারি কার্যবিধি ওই বিশেষ আইনের পদ্ধতি বাতিল করতে পারবে না — বিপরীত কোনো সুনির্দিষ্ট বিধান না থাকলে বিশেষ আইনের পদ্ধতি অক্ষুণ্ণ থাকবে।",
  "NO_IRAC_Reasoning": "The Code of Criminal Procedure, 1898, Section 1(2) অনুযায়ী, ফৌজদারি কার্যবিধি সারা দেশে প্রযোজ্য হলেও অন্য বিশেষ আইনের নিজস্ব পদ্ধতি থাকলে সেটি বহাল থাকে। শুধুমাত্র কার্যবিধিতে সুনির্দিষ্ট বিপরীত বিধান থাকলে বিশেষ আইনের পদ্ধতি বাতিল হতে পারে।",
  "Type": "Non-Obstacle Clause",
  "Difficulty": "Medium",
  "Keywords": [
    "বিশেষ আইন",
    "প্রত্যয়র",
    "প্রয়োগযোগ্যতা",
    "বিশেষ পদ্ধতি",
    "সরক্ষণ ধারা"
  ],
  "Cited Acts and Sections": "The Code of Criminal Procedure, 1898, Section 1(2)"
}

{
  "Section Number": "Section 32",
  "Entry_ID": 2,
  "Question": "Scenario-based: একজন সাক্ষীকে আদালতে হাজির হতে সমন দেওয়া হয়েছে, কিন্তু তিনি আসেননি। আদালত সর্বোচ্চ কত টাকা জরিমানা করতে পারে, এবং তিনি যদি ভবিষ্যতে হাজির থাকার নিশ্চয়তা দিতে অস্বীকার করেন তাহলে কী হবে?",
  "Subsection/Clause": "Clauses (c) and (d)",
  "Section Text": "The Court may compel the attendance of any person to whom a summons has been issued under section 30 and for that purpose may- (a) issue a warrant for his arrest; (b) attach and sell his property; (c) impose a fine upon him not exceeding five hundred Taka; (d) order him to furnish security for his appearance and in default commit him to the civil prison.",
  "IRAC_Reasoning": {
    "Issue": "সমন-প্রাপ্ত সাক্ষী অনুপস্থিত থাকলে আদালত সর্বোচ্চ কত জরিমানা ও পরবর্তী শাস্তি আরোপ করতে পারে?",
    "Rule": "অনুপস্থিতির ক্ষেত্রে পর্যায়ক্রমিক শাস্তি প্রযোজ্য — প্রথম পচিশত টাকা পর্যন্ত জরিমানা, তারপর ভবিষ্যতে হাজির না হওয়ায় নিশ্চয়তা না দিলে দেওয়ানি কারাগার।",
    "Application": "Clause (c) আর্থিক সীমা পচিশত টাকা নির্ধারণ করে, Clause (d) পরবর্তীতে কার্যকর হয় — নিশ্চয়তা দিতে বার্থ হলে দেওয়ানি কারাগার বাধ্যতামূলক। দুটি clause বিকল্প নয়, ক্রমানুসারে প্রযোজ্য।",
  },
  "Answer": "The Code of Civil Procedure, 1908, Section 32, Clauses (c) and (d) অনুযায়ী, আদালত সর্বোচ্চ পচিশত টাকা জরিমানা করতে পারে এবং সাক্ষী যদি ভবিষ্যতে হাজির থাকার নিশ্চয়তা দিতে অস্বীকার করেন তাহলে দেওয়ানি কারাগারে পাঠাতে পারে।",
  "NO_IRAC_Reasoning": "The Code of Civil Procedure, 1908, Section 32, Clauses (c) and (d) অনুযায়ী, আদালতের সমন উপেক্ষা করলে সাক্ষীকে সর্বোচ্চ পচিশত টাকা জরিমানা করা যায়। এরপরও যদি তিনি ভবিষ্যতে হাজির থাকার নিশ্চয়তা দিতে অস্বীকার করেন, তাহলে আদালত তাকে দেওয়ানি কারাগারে পাঠাতে পারে।",
  "Type": "Procedural Chronology",
  "Difficulty": "Medium",
  "Keywords": [
    "সমন",
    "জরিমানা",
    "দেওয়ানি কারাগার",
    "হাজিরার নিশ্চয়তা",
    "ব্যাহতকরণ"
  ],
  "Cited Acts and Sections": "The Code of Civil Procedure, 1908, Section 32(c) and Section 32(d)"
}
```

Figure 5: Bangla single-hop generation prompt, part 2 of 2, reproduced from the original report.

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 designed to train a model on "Relevant Section Selection." The model must use three-step reasoning (Issue, Rule, Application) to eliminate distractors and identify the governing section. All IRAC_Reasoning values must be written in Bangla. Answer is the direct output after Application — there is no Conclusion field.

[SFT Data Quality Context]

You are generating supervised fine-tuning (SFT) data for smaller language models (1.6B-8B parameters). Your outputs are the training signal — not a demonstration of your own capability. For this task, the critical signal is the elimination reasoning: each distractor must be rejected with a precise, one-sentence legal reason in Bangla. Smaller models learn section-selection by internalizing the pattern of substantive one-sentence eliminations repeated consistently across entries — they cannot extract that pattern from verbose, uneven, or structurally inconsistent rejections. Vague rejections teach nothing, padding dilutes the pattern. Sentence count and format consistency across entries is as important as legal accuracy. Every constraint in this prompt exists to produce data the model can generalize from, not data that reads impressively in isolation.

[Jurisdiction & Corpus]

Jurisdiction : [Bangladesh / India / Pakistan / UK] — select one.

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

Citation Note : Legislative addresses vary by instrument. Use whichever structure the provision actually uses — a Section, an Order and Rule, a Schedule and Article, or a combination. Never invent a Section number if the provision is an Order, Rule, or Article.

[Scope of Work]

Target Range: Generate entries for Sections [Start Section] to [End Section].

Volume: Provide [number of questions] questions.

[Output Style — SFT Quality Targets]

These targets apply to every entry. They exist to produce concise, consistent, high-signal SFT data.

- IRAC_Reasoning Issue: সর্বোচ্চ ১২ শব্দ।
- IRAC_Reasoning Rule object: প্রতিটি section-এর জন্য ঠিক এক বাক্য, কোনো ব্যতিক্রম নেই।
- IRAC_Reasoning Application object: প্রতিটি section-এর জন্য ঠিক এক বাক্য, কোনো ব্যতিক্রম নেই।
- Answer field: ২৫-৪৫ শব্দ পর্য্যে।
- অন্তঃসেজনীয় বিভাগ এড়িয়ে চলুন। প্রতিটি শব্দ প্রয়োজনীয় হতে হবে।

[Format & Schema]

Output must be a strictly valid JSON array of objects.

Schema per object:

Question : বাংলা স্পষ্ট, স্বাধীন প্রশ্ন (উচ্চ কঠিনতার জন্য "Scenario-based" ব্যবহার করুন)। প্রশ্নটি সাধারণ মানুষের ভাষায় (client-speak) লিখতে হবে — একজন মডেল যেভাবে তার সমস্যা বর্ণনা করেন সেভাবে। আইনি পরিভাষা IRAC_Reasoning এবং Answer-এ ব্যবহার করতে হবে, প্রশ্নে নয়।

Entry ID : চর্চামূলক generation run-এর মধ্যে entry-র একটি কৃত্রিম পূর্ণসংখ্যা শনাক্তকারী (যেমন, 1, 2, 3)। একটি নির্দিষ্ট key নাম ব্যবহার করুন — সংখ্যাটি key-এর মধ্যে এম্বেড করবেন না।

Section Number : সঠিক governing provision-এর legislative address।

Subsection/Clause : The specific sub-provision or clause the question focuses on.

Section Text : The full, verbatim text of the correct section from the dataset (whole section, not only the clause or subsection).

Possible Sections : ঠিক ৪টি object-এর তালিকা, প্রতিটিতে Section Number এবং Full Text (সঠিক section + ৪টি distractor)। Full text অবশ্যই verbatim কপি করতে হবে। সঠিক section-এর অবস্থান — প্রশ্ন, দ্বিতীয়, তৃতীয়, চতুর্থ বা পঞ্চম — generation batch জুড়ে

randomly পরিবর্তন করুন, যা যা সমান (স্কোরেয়ার্থিত) স্বাক্ষরায় স্থানীয় স্থানীয় না।

IRAC_Reasoning : সঙ্গত মান বাংলায়। তিনটি key সহ একটি object — Issue, Rule, Application। Conclusion field নেই — Answer হলো Application-এর পরের সরাসরি output। শুধুমাত্র অভ্যন্তরীণ মান-নিয়ন্ত্রণের জন্য। কোনো chatbot বা UI-তে দেখানো যাবে না।

Answer : ব্যবহারকারীর জন্য একমাত্র দৃশ্যমান উত্তর। নিচের Answer Field নিয়মাবলী দেখুন।

NO_IRAC_Reasoning : IRAC কঠোর অভ্যাসের জন্য নিচের NO_IRAC_Reasoning Field নিয়মাবলী দেখুন।

Relevant Section : প্রশ্নটির উত্তর দেয় এমন একমাত্র সঠিক section number।

Type : নিচের Type Selection Method অনুসরণ করে একটি type নির্বাচন করুন।

Difficulty : "Very High (Contextual Selection)"।

Keywords : 3-5 legal terms or concepts central to the section, in Bangla.

Cited Acts and Sections : The primary Act name (in English) and the specific legislative address.

[IRAC_Reasoning Keys — সঙ্গত মান বাংলায়]

Note: IRAC_Reasoning হলো Answer-এ পৌঁছানোর অভ্যন্তরীণ reasoning chain। মান-নিয়ন্ত্রণের জন্য JSON-এ সংরক্ষিত থাকবে, কিন্তু সকল UI/chatbot-এ লুকিয়ে রাখতে হবে।

"Issue" : এক বাক্য। পরিস্থিতিতে উপস্থিত সুনির্দিষ্ট আইনি প্রশ্নটি বর্ণনা। ঘটনার বিবরণ বাদ দিয়ে শুধু মূল আইনি সমস্যাটি চিহ্নিত করুন।

"Rule" : প্রতিটি Possible Section-এর জন্য একটি করে key সহ একটি object। প্রতিটি ৫টি প্রার্থী section-এর জন্য ঠিক এক বাক্য। যে section-টি কী নিয়ন্ত্রণ করে তা পার্যাক্সেজ করুন। (কোনো section-এর পাঠ পুনরুৎপাদন করবেন না। এক বাক্য, কোনো ব্যতিক্রম নেই।)

"Application" : প্রতিটি Possible Section-এর জন্য একটি করে key সহ একটি object। ৪টি distractor section-এর প্রত্যেকটির জন্য ঠিক এক বাক্য। স্পষ্টভাবে ব্যাখ্যা করুন কেন একটি প্রশ্নটির উত্তর নিতে পারে না — scope mismatch, ভিন্ন trigger condition, বা বিষয়বস্তুর পার্থক্যে চিহ্নিত। সঠিক section-এর জন্য এক বাক্য। ব্যাখ্যা করুন কেন এটি সরাসরি ও একত্রোচ্চভাবে প্রশ্নটির উত্তর দেয়। শুধু দাবি না, আইনি কারণ দেখান। এক বাক্য, কোনো ব্যতিক্রম নেই। প্রতিটি entry-তে elimination ব্যাক্তার শুরুতর গঠন পরিবর্তন করুন — একই phrase বা বাক্য-গঠন প্রতিটি entry-র distractor rejection শুরু করতে হবে না।

বিশেষ দৃষ্টব্য: Application-এর পরের reasoning chain সম্পূর্ণ। Answer হলো plain-language output। JSON-এ Conclusion key যোগ করবেন না।

[Answer Field — ব্যবহারকারীর জন্য একমাত্র দৃশ্যমান উত্তর]

Answer কেবল IRAC_Reasoning-এর ঠিক পরে স্থাপন করতে হবে।

নিয়মাবলী:

- Act-এর নাম (ইংরেজিতে) দিয়ে শুরু করুন, তারপর সঠিক legislative address এবং "অনুযায়ী"। যেমন: Section: "The Evidence Act, 1872, Section 32, Clause (1) অনুযায়ী, ..."
- Order and Rule: "The Code of Civil Procedure, 1908, Order XXXI, Rule, 1 অনুযায়ী, ..."
- Schedule/Article: "The Limitation Act, 1908, Schedule I, Article 23 অনুযায়ী, ..."
- Combined: Act-এ যে ক্রমে আছে সেই ক্রমে সকল প্রয়োজ্য স্তর (Subsection / Clause / Illustration / Rule) অন্তর্ভুক্ত করুন।
- "সঠিক ধরা হলো" : বা এই ধরনের meta-language ব্যবহার করবেন না — সরাসরি প্রশ্নের উত্তর দিন।
- সহজ, সরল বাংলায় — প্রশ্নটি যেভাবে জিজ্ঞেস করা হয়েছে সেভাবে উত্তর দিন।
- সর্বোচ্চ এক থেকে দুই বাক্য। ২৫-৪৫ শব্দ।
- উদ্ভূতজন জন প্রয়োজনীয় পরিভাষার বাইরে আইনি জটিল ভাষা পরিহার করুন।

উদাহরণ:

"Answer": "The Code of Civil Procedure, 1908, Section 16, Clause (f) অনুযায়ী, মামলাটি সেই আদালতে দায়ের করতে হবে যার স্থানীয় এখতিয়ারের মধ্যে হীরার আটটি রয়েছে — আদালত-সংযুক্ত চলমান সম্পত্তি উদ্ধারের মামলায় সম্পত্তির অবস্থানস্থলের আদালতই স্বাভাবিকতম প্রতিদায়নপ্রাপ্ত।"

[NO_IRAC_Reasoning Field — ব্যবহারকারীর জন্য নয়, training signal হিসেবে]

Answer-এর ঠিক পরে স্থাপন করতে হবে।

নিয়মাবলী:

- দুই থেকে তিন বাক্য সহজ বাংলায় — Issue / Rule / Application / Conclusion লেবেল ব্যবহার করা যাবে না।
- সম্পূর্ণ legislative citation অবশ্যই prose-এর মধ্যে স্বাভাবিকভাবে অন্তর্ভুক্ত করতে হবে — Act-এর নাম এবং legislative address। Citation বাদ দেওয়া যাবে না। IRAC_Reasoning এবং NO_IRAC_Reasoning-এর মধ্যে একমাত্র পার্থক্য হলো কাঠামোবদ্ধ সেবনের উপস্থিতি বা অনুপস্থিতি, citation-এর নয়।
- সঠিক section কোন প্রয়োজ্য এবং সবচেয়ে কাছে distractor কোন প্রয়োজ্য নয় — উভয়ই সহজ ভাষায় ব্যাখ্যা করুন।
- Section Text ছব্ব পুনরুৎপাদন করবেন না। শুধু পার্যাক্সেজ করুন।
- লক্ষ্যমাত্রা: ৪০-৬০ শব্দ।

উদাহরণ:

"NO_IRAC_Reasoning": "The Code of Civil Procedure, 1908, Section 16, Clause (f) অনুযায়ী, এই মামলাটি ক্ষতিপূরণের নয় বরং আটটি সরাসরি ক্ষেত্রত পাওয়ার মামলা — এই পার্থক্যই Section 19 ব্যতিল করে এবং Section 16(f)-কে প্রয়োজ্য করে। Section 16(f) আদালত-সংযুক্ত চলমান সম্পত্তি উদ্ধারের মামলা সরাসরি নিয়ন্ত্রণ করে এবং সম্পত্তি যেখানে আছে সেই আদালতে প্রতিদায়ন নির্ধারণ করে।"

[Type Selection Method]

সহায়ক ফাইল QUESTION_TYPE_VOCABULARY.md দেখুন।

প্রতিটি প্রশ্ন তৈরির সময় নিম্নলিখিত প্রক্রিয়া অনুসরণ করুন:

- Step 1 — QUESTION TYPE VOCABULARY.md স্বাধীনতার শেষে Quick Reference তালিকা থেকে একটি type এনে মেশোলাভাবে নির্বাচন করুন।
- Step 2 — "যাচাই করুন" যে নির্বাচিত type-টি প্রক্রিয়াদ্বারা ধারায় প্রকৃতপক্ষে পরীক্ষা করা যায়। যদি হয়, সেটি ব্যবহার করুন। যদি না হয়, পরবর্তী সবচেয়ে উপযুক্ত type নির্বাচন করুন।
- Step 3 — প্রথম প্রশ্ন দিয়ে পরে type নির্বাচন করবেন না — বরং type নির্বাচন করে তারপর সেই type-এর উপযোগী প্রশ্ন ও পরিচ্ছিন্ন তৈরি করুন।
- Step 4 — যেকোনো ১০টি প্রশ্নের ব্যাটে কমপক্ষে ৩টি ভিন্ন type নিশ্চিত করুন। একই type পরপর দুইবারের বেশি ব্যবহার করবেন না।

[Distractor Selection Logic]

নিম্নলিখিত প্রক্রিয়ায় ৪টি distractor section নির্বাচন করুন:

- Step 1 — সঠিক section-এর মূল আইনি ধাপাগুলো চিহ্নিত করুন: এটি কী নিয়ন্ত্রণ করে? কী মূল পরিভাষা ব্যবহার করে? কোন পদ্ধতিগত পর্যায়ে প্রযোজ্য?
- Step 2 — ডেটাসেটের অন্য সকল section-কে নিম্নলিখিত মানদণ্ডে মূল্যায়ন করুন এবং সর্বোচ্চ সামগ্রিক মিল আছে এমন ৪টি নির্বাচন করুন:
 - (ক) একই বা সঙ্গত আইনি বিষয়বস্তু
 - (খ) একই বা ওভারল্যাপিং মূল পরিভাষা
 - (গ) একই বা অনুরূপ পদ্ধতিগত প্রেক্ষাপটে প্রযোজ্য
 - (ঘ) প্রথম দুটিতে সঙ্গত উত্তর মনে হয় কিন্তু গভীর বিশ্লেষণে scope সীমাবদ্ধতা, ভিন্ন trigger condition বা সংকীর্ণ/বিস্তৃত প্রয়োগের কারণে ব্যতিল হয়
- Step 3 — যাচাই করুন যে নির্বাচিত ৪টি distractor-এর কোনোটিই প্রশ্নটির উত্তর দেয় না, কিন্তু প্রতিটি সূক্ষ্ম বিশ্লেষণের আগে একজন যুক্তিসঙ্গত পাঠকের কাছে সঙ্গত বা উত্তর হিসেবে মনে হতে পারে।
- Step 4 — Possible Sections-এ সঠিক section-এর অবস্থান randomly পরিবর্তন করুন। Generation batch জুড়ে প্রতিটি অবস্থানে (১ম থেকে ৫ম) প্রায় সমান ফ্রিকোয়েন্সিতে রাখুন। সবসময় প্রথমে রাখবেন না।

[Constraint Guidelines]

Verbatim Integrity : Possible Sections-এর সমস্ত "Full Text" এবং "Section Text" JSON ডেটাসেট থেকে ছব্ব কপি করতে হবে। খুঁজি থেকে পুনর্নির্মাণ বা সংক্ষিপ্ত করা যাবে না। যদি সঠিক পাঠ পাওয়া না যায়, entry-টি incomplete হিসেবে চিহ্নিত করুন।

Anti-Hallucination : খুঁজি থেকে কখনো section-এর পাঠ পুনর্নির্মাণ বা পার্যাক্সেজ করবেন না।

Client-Speak Rule : প্রশ্নটি অবশ্যই একজন সাধারণ মডেলের ভাষায় হতে হবে — প্রশ্নের মধ্যে আইনি পরিভাষা ব্যবহার করবেন না।

Structural Order : উপরে উল্লিখিত ক্রমে JSON keys রাখুন।

IRAC Rule Step : শুধু প্যারাফ্রেজ করুন। IRAC Reasoning-এর ভেতরে section-এর পাঠ পুনরুৎপাদন করবেন না।

IRAC Application : Distractor elimination অবশ্যই অবিলম্বে যুক্তিসংগত হতে হবে — শুধু দাবি নয়, scope mismatch বা trigger condition-এর পার্থক্য যাচ্যা করুন। যে distractor-টি সঠিক section-এর যত বেশি কাছাকাছি, তার elimination তত বেশি সুনির্দিষ্ট হতে হবে।

Length Discipline : কোনো field তার শব্দসীমা অতিক্রম করলে আউটপুটের আগে সংশোধন করুন।

Answer Visibility Rule: IRAC Reasoning কোনো ব্যবহারকারীমুখী interface বা chatbot-এ প্রদর্শন করা হবে না।

Scenario Freshness : পদ্যের entry-তে একই চরিত্রের নাম, চেনানোদের ধরন বা বিবরণের পরিবর্তন পুনরায় ব্যবহার করবেন না।

Multi-Clause Sections : সঠিক section-এ একাধিক clause থাকলে, প্রদ্রাটি যে নির্দিষ্ট clause নিয়ন্ত্রণ করে তা Subsection/Clause field-এ উল্লেখ করুন। একাধিক clause একসাথে উত্তর দিতে পারে এমন প্রশ্ন তৈরি করবেন না।

[Pre-Output Quality Check — আউটপুট দেওয়ার আগে প্রতিটি এন্ট্রির জন্য নিচের সকল বিষয় যাচাই করুন]

- Possible Sections-এর সকল "Full Text" — ডেস্টিনেশি থেকে বের করি করা হয়েছে
- "Section Text" — ডেস্টিনেশি থেকে ছবছ করি করা হয়েছে
- Possible Sections-এ সঠিক section-এর অবস্থান — randomly পরিবর্তিত, সবসময় প্রথম নয়
- Rule object — প্রতিটি entry টিক এক হোক। প্যারাফ্রেজ করা; কোনো section পাঠ সরাসরি করি করা হয়নি
- Application object — প্রতিটি entry টিক এক হোক। প্রতি distractor-এর জন্য অবিলম্বে যুক্তিসংগত elimination কারণ; সঠিক section-এর জন্য সরাসরি প্রয়োজ্যতার কারণ
- "Answer" — Act নাম দিয়ে শুরু হয়েছে; meta-language নেই; ২৫-৪৫ শব্দ; সহজ বাংলায়
- "NO_IRAC_Reasoning" — দুই থেকে তিন বাক্য; IRAC স্টেপে নেই; Act নাম ও legislative address prose-এ অন্তর্ভুক্ত; সঠিক selection এবং মূল distractor distinction বাখ্যা করা হয়েছে; ৪০-৬০ শব্দ
- "Type" — QUESTION-TYPE VOCABULARY-এর ভেতরে নির্ভরন করা হয়েছে; উদ্ভাবিত বা ফাঁকা নয়
- সময় ভাষা — সাধারণ মন্তব্যের ভাষায়, অর্থাৎ পরিত্রাযুক্ত
- Distractor বৈচিত্র্য (Step 3) — set-এ কমপক্ষে একটি same-chapter distractor, একটি scope-mismatch distractor, এবং একটি common-misreading distractor আছে

[Negative Example — What to Avoid]

The following shows a FAILING IRAC_Reasoning for an advanced selection entry (Bangla). Errors annotated. Do not produce entries like this.

```
"IRAC_Reasoning": {
  "Issue": "সকল ধারাটি প্রযোজ্য যখন বাদী সংযুক্ত সম্পত্তি ফেরত পেতে চান এবং সেই সম্পত্তির বিরুদ্ধে বিবাদী অনায় করছেন বলে দাবি করেন?",
  // X FAIL — Issue: অনেক দীর্ঘ এবং দুটি পরস্পর-বিরোধী fact pattern একসাথে আছে। সর্বোচ্চ ১৫ শব্দ।

  "Rule": {
    "Section 16": "suits for the recovery of movable property actually under distraint or attachment, shall be instituted in the Court within the local limits of whose jurisdiction the property is situate.",
    // X FAIL — Rule: Section Text ছবছ উদ্ভূত করা হয়েছে। নিজের ভাষায় paraphrase করতে হবে।
    "Section 19": "এই ধারাটি এখানে প্রযোজ্য নয়।",
    // X FAIL — Rule: এটি Application-এর বিফল, Rule নয়। Rule অবশ্যই বলতে হবে ধারাটি কী নিয়ন্ত্রণ করে।
  },
  "Application": {
    "Section 19": "প্রযোজ্য নয়।",
    // X FAIL — Application: Bare assertion। সুনির্দিষ্ট scope mismatch চিহ্নিত করতে হবে — Section 19-এর trigger condition (অন্যায়ের অতিপূরণ) বাদীর দাবি (সংযুক্ত সম্পত্তি উদ্ধার) থেকে কীভাবে আলাদা তা এক বাক্যে বলতে হবে।
  }
}
```

সংশোধিত সংস্করণ নিচের Example-এ দেখুন।

```
Example:
{
  "Question": "Scenario-based: একটি আদালতের আদেশ করে যীরার আটটি জন্ম করা হয়েছে। মালিক এখন সেই আটটি ফেরত পেতে মামলা করেন। কোন আদালতে মামলা দায়ের করতে হবে?",
  "Entry_ID": 1,
  "Section Number": "Section 16",
  "Subsection/Clause": "Clause (f)",
  "Section Text": "Subject to the pecuniary or other limitations prescribed by any law, suits- (a) for the recovery of immovable property with or without rent or profits, (b) for the partition of immovable property, (c) for foreclosure, sale or redemption in the case of a mortgage of or charge upon immovable property, (d) for the determination of any other right to or interest in immovable property, (e) for compensation for wrong to immovable property, (f) for the recovery of movable property actually under distraint or attachment, shall be instituted in the Court within the local limits of whose jurisdiction the property is situate...",
  "Possible Sections": {
    {
      "Section Number": "Section 19",
      "Full Text": "Where a suit is for compensation for wrong done to the person or to movable property, if the wrong was done within the local limits of the jurisdiction of one Court and the defendant resides... the suit may be instituted at the option of the plaintiff in either of the said Courts."
    },
    {
      "Section Number": "Section 20",
      "Full Text": "Subject to the limitations aforesaid, every suit shall be instituted in a Court within the local limits of whose jurisdiction- (a) the defendant... at the time of the commencement of the suit, actually or voluntarily resides, or carries on business, or personally works for gain..."
    },
    {
      "Section Number": "Section 16",
      "Full Text": "Subject to the pecuniary or other limitations prescribed by any law, suits- (a) for the recovery of immovable property... (f) for the recovery of movable property actually under distraint or attachment, shall be instituted in the Court within the local limits of whose jurisdiction the property is situate..."
    },
    {
      "Section Number": "Section 17",
      "Full Text": "Where a suit is to obtain relief respecting, or compensation for wrong to, immovable property situate within the jurisdiction of different Courts, the suit may be instituted in any Court within the local limits of whose jurisdiction any portion of the property is situate..."
    },
    {
      "Section Number": "Section 15",
      "Full Text": "Every suit shall be instituted in the Court of the lowest grade competent to try it."
    }
  },
  "IRAC_Reasoning": {
    "Issue": "আদালত-সংযুক্ত চলমান সম্পত্তি উদ্ধারের মামলা কোন আদালতে দায়ের করতে হবে?",
    "Rule": {
      "Section 19": "কোনো ব্যক্তি বা চলমান সম্পত্তির বিরুদ্ধে অন্যায়ের অতিপূরণের মামলায় বাদীকে অন্যায়ের স্থান বা বিবাদীর বাসস্থানের আদালতের মধ্যে যেকোনো একটি বেছে নেওয়ার অধিকার দেয়।",
      "Section 20": "অবশ্যই আদালত প্রযোজ্যের ভাষায় বিধি নিয়ন্ত্রণ করে — বিবাদীর বাসস্থান বা মামলার কারণ উদ্ধারের স্থান মামলা করার অনুমতি দেয়, তবে অধিক সুনির্দিষ্ট বিধানের কাছে নতি স্বীকার করে।",
      "Section 16": "স্থাবর সম্পত্তি সংক্রান্ত মামলার এখতিয়ার নিয়ন্ত্রণ করে এবং সুনির্দিষ্টভাবে আদালত-সংযুক্ত চলমান সম্পত্তি উদ্ধারের মামলায় সম্পত্তি যেখানে আছে সেই আদালতকে এখতিয়ার দেয়।",
      "Section 17": "একাধিক আদালতে এখতিয়ার বিস্তৃত স্থাবর সম্পত্তি সংক্রান্ত মামলায় যেকোনো অংশের অধিকারের আদালতে দায়ের করার সুযোগ দেয়।",
      "Section 15": "মামলার আদালত স্থান নয়, বরং কোন শ্রেণীর আদালত মামলাটি পরিচালনা করতে সক্ষম সেই বিষয়টি নিয়ন্ত্রণ করে।"
    },
    "Application": {
      "Section 19": "প্রযোজ্য নয় — এই ধারা শুধুমাত্র চলমান সম্পত্তির বিরুদ্ধে অন্যায়ের অতিপূরণের মামলায় প্রযোজ্য; এখানে বাদী অতিপূরণ নয়, বরং আটটি সরাসরি ফেরত চাইছেন, তাই এটি উদ্ধারের মামলা, অতিপূরণের মামলা।",
      "Section 20": "প্রযোজ্য নয় — এটি লম্বিত অংশিত বিধান বা অধিক সুনির্দিষ্ট ধারার কাছে নতি স্বীকার করে; Section 16(f) সরাসরি প্রযোজ্য হওয়ায় এটি হান্ধাত্ত হয়।",
      "Section 16": "সরাসরি প্রযোজ্য — যীরার আটটি আদালত-সংযুক্ত চলমান সম্পত্তি, যা Clause (f)-এ সুনির্দিষ্ট trigger condition, এবং এখতিয়ার ব্যাখ্যামূলকভাবে সম্পত্তি যেখানে আছে সেই আদালতে নির্ধারিত।",
      "Section 17": "প্রযোজ্য নয় — এই ধারা একচেটিয়াভাবে একাধিক এখতিয়ার বিস্তৃত স্থাবর সম্পত্তির ক্ষেত্রে প্রযোজ্য; যীরার আটটি চলমান সম্পত্তি, তাই Section 17-এর trigger condition পূরণ হয় না।",
      "Section 15": "প্রযোজ্য নয় — প্রদ্রা কোন আদালতের অফসে মামলা করতে হবে তা নিয়ে, কোন শ্রেণীর আদালত উপযুক্ত তা নিয়ে নয় — এই দুটি ভিন্ন এখতিয়ার-প্রদ্রা।"
    }
  },
  "Answer": "The Code of Civil Procedure, 1908, Section 16, Clause (f) অনুযায়ী, মামলাটি সেই আদালতে দায়ের করতে হবে যার স্থানীয় এখতিয়ারের মধ্যে যীরার আটটি রয়েছে — আদালত-সংযুক্ত চলমান সম্পত্তি উদ্ধারের মামলায় সম্পত্তির অবস্থানস্থলের আদালতই ব্যাখ্যামূলক এখতিয়ারপ্রাপ্ত।",
  "NO_IRAC_Reasoning": "The Code of Civil Procedure, 1908, Section 16, Clause (f) অনুযায়ী, এই মামলাটি অতিপূরণের নয় বরং আটটি সরাসরি ফেরত পাওয়ার মামলা — এই পার্থক্যই Section 19 ব্যক্তি করে এবং Section 16(f)-কে প্রযোজ্য করে। Section 16(f) আদালত-সংযুক্ত চলমান সম্পত্তি উদ্ধারের মামলা সরাসরি নিয়ন্ত্রণ করে এবং সম্পত্তি যেখানে আছে সেই আদালতে এখতিয়ার নির্ধারণ করে।",
  "Relevant Section": "Section 16",
  "Type": "Jurisdictional / Territorial",
  "Difficulty": "Very High (Contextual Selection)",
  "Keywords": {
    "চলমান সম্পত্তি",
    "আটক",
    "সংযুক্ত",
    "স্থানীয় সীমা",
    "উদ্ধার মামলা"
  },
  "Cited Acts and Sections": "The Code of Civil Procedure, 1908, Section 16(f)"
}
```

Figure 6: Bangla advanced-selection generation prompt, part 2 of 2, reproduced from the original report.

Role: Expert Legal Educator and Bar Examination Specialist.
Tag: তোমাকে সরবরাহ করা হবে পরীক্ষার JSON ডেটাসেট থেকে প্রশ্ন নিয়ে হবে — প্রতিটি প্রশ্ন স্বনির্ভরশীল উত্তর (ক, খ, গ, ঘ) এবং সঠিক উত্তর দেওয়া আছে। প্রতিটি প্রশ্নের জন্য সরবরাহ করা আইনি ডেটাসেট থেকে নিয়ন্ত্রণকারী ধারাটি চিহ্নিত করে। IRAC পদ্ধতিতে স্রষ্টা ধারাগুলো বিবেচনা করে সঠিক MCQ উত্তর নির্বাচন করে, সঠিক ধারার উদ্ধৃতি দিয়ে উত্তর দাও এবং Answer-এর সর্বশেষ থাকে সঠিক বিকল্পের অক্ষর উল্লেখ করে। All IRAC_Reasoning values must be written in Bangla.

[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 — not a demonstration of your own capability. For this task, the critical training signal is two-layered: the model must learn (1) section-selection via one-sentence eliminations, and (2) MCQ resolution — deriving which lettered option follows from the governing provision. Smaller models acquire both skills through dense repetition of the same tight pattern: one sentence per section in Rule, one sentence per section in Application, a clean Answer ending with the correct letter. Verbose explanations dilute both signals. Inconsistent Answer endings prevent the model from learning the MCQ-resolution pattern. Sentence count and format consistency across entries is as important as legal accuracy. Every constraint in this prompt exists to produce data the model can generalise from, not data that reads impressively in isolation.

[Jurisdiction & Corpus]
Jurisdiction : Bangladesh.
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.
Citation Note : Legislative addresses vary by instrument. Use whichever structure the provision actually uses — a Section, an Order and Rule, a Schedule and Article, or a combination. Never invent a Section number if the provision is an Order, Rule, or Article.

[Input Sources]
Bar Exam Dataset : সরবরাহ করা হবে পরীক্ষার JSON — প্রতিটি object-এ রয়েছে: question_no, question_text, options (ক/খ/গ/ঘ), correct_answer, correct_answer_text।
Law Dataset : সরবরাহ করা আইনি JSON ডেটাসেট যাতে ধারার পাঠ রয়েছে (যেমন: The Code of Civil Procedure, The Evidence Act ইত্যাদি) — শুধুমাত্র Possible Sections এবং Section Text-এর জন্য ছব্ব ধারার পাঠ নেওয়ার কাজে ব্যবহৃত হবে।

[Scope of Work]
Target Range: বার পরীক্ষার ডেটাসেট থেকে [Start Question No.] থেকে [End Question No.] পর্যন্ত প্রশ্নগুলো প্রক্রিয়া করা।
Volume: নির্ধারিত পত্রিকার প্রতিটি পত্র পরীক্ষার প্রশ্নের জন্য একটি করে entry। বাতিল বা বিকল্পিত প্রশ্ন বাদ দেবে না — পরিবর্তে Cancelled field ব্যবহার করে।

[Output Style — SFT Quality Targets]
These targets apply to every entry. They exist to produce concise, consistent, high-signal SFT data.
- IRAC_Reasoning Issue: সর্বোচ্চ ১৫০ শব্দ।
- IRAC_Reasoning Rule object: প্রতিটি section-এর জন্য ঠিক এক বাক্য, কোনো ব্যতিক্রম নেই।
- IRAC_Reasoning Application object: প্রতিটি section-এর জন্য ঠিক এক বাক্য, কোনো ব্যতিক্রম নেই।
- Answer field: সর্বোচ্চ ৩৫-৬০ শব্দ (MCQ অক্ষর-উপসংহার অন্তর্ভুক্ত)।
- অপ্রয়োজনীয় বিস্তার এড়িয়ে চলুন। প্রতিটি শব্দ প্রয়োজনীয় হতে হবে।

[Format & Schema]
Output must be a strictly valid JSON array of objects.
Schema per object:
Question : বার পরীক্ষার JSON-এর question_text ছব্ব কপি। পরিবর্তন বা সংশোধন করা যাবে না।
Entry_ID : চতুর্থমা generation run-এর যথা entry-র একটি ক্রমিক পূর্ণসংখ্যা শনাক্তকারী (যেমন, 1, 2, 3)। একটি নির্দিষ্ট key নাম ব্যবহার করুন — সংঘাতী key-এর মধ্যে এষণ্ড করবেন না।
Question_No : বার পরীক্ষার JSON-এর question_no মান (যেমন, "১", "২")।
Bar_Exam_Options : বার পরীক্ষার JSON-এর options object — ক, খ, গ, ঘ keys সহ তাদের ছব্ব পাঠ। অবিকল কপি করতে হবে।
Correct_Option : সঠিক উত্তরকে একটি মাত্র অক্ষর (যেমন, "ক", "খ", "গ", "ঘ")। বার পরীক্ষার JSON-এর correct_answer থেকে নেওয়া।
Section_Number : নিয়ন্ত্রণকারী ধারার legislative address। কোনো একক ধারা সরাসরি প্রযোজ্য না হলে 'General Principle' ব্যবহার করে।
Subsection/Clause : সঠিক উত্তরের সাথে প্রাসঙ্গিক নির্দিষ্ট উপ-ধারা বা clause। প্রযোজ্য না হলে 'N/A' বা 'General Provision' ব্যবহার করে।
Section_Text : আইনি ডেটাসেট থেকে নিয়ন্ত্রণকারী ধারার সম্পূর্ণ ছব্ব পাঠ। Section Number যদি 'General Principle' হয় তাহলে 'N/A' — No single provision governs directly.' ব্যবহার করে।
Possible_Sections : প্রতিটি object-এর বালিস, প্রতিটিতে Section Number এবং Full Text (নিয়ন্ত্রণকারী ধারা + গঠিত distractor, অথবা কোনো সরাসরি ধারা না থাকলে এটি সঙ্গত্য প্রার্থী)। Full text অবশ্যই আইনি ডেটাসেট থেকে ছব্ব কপি করতে হবে। নিয়ন্ত্রণকারী ধারার অবশ্যই — প্রথম, দ্বিতীয়, তৃতীয়, চতুর্থ বা পঞ্চম — generation batch জুড়ে randomly পরিবর্তন করুন, প্রায় সমান হ্রিকোয়েলিডিতে। সবসময় প্রথমে রাখবেন না।
IRAC_Reasoning : সকল মান বালিস। চিহ্নিত key সহ একটি object — Issue, Rule, Application। Conclusion field নেই — Answer হলো Application-এর পরের সরাসরি output। শুধুমাত্র অভ্যন্তরীণ মান-নিয়ন্ত্রণের জন্য। কোনো chatbot বা UI-তে দেখানো যাবে না।
Answer : ব্যবহারকারীর জন্য একমাত্র প্রশামন উত্তর। নিচের Answer Field নিয়মাবলী দেখুন।
NO_IRAC_Reasoning : IRAC ইচ্ছামে ছাড়া সহজ বাধ্য। নিচের NO_IRAC_Reasoning Field নিয়মাবলী দেখুন।
Relevant_Section : নিয়ন্ত্রণকারী ধারার legislative address, অথবা Possible Sections-এ সরাসরি নিয়ন্ত্রণকারী কোনো ধারা না থাকলে 'None'।
Cancelled : Boolean। বার পরীক্ষার JSON-এ correct_answer যদি 'বাতিল' বা সমতুল্য হয় তাহলে true, অন্যথায় false। true হলে Relevant_Section-কে 'N/A' এবং Answer-কে 'এই প্রশ্নটি মূল পরীক্ষার বাতিল করা হয়েছিল।' সেট করে এবং IRAC_Reasoning-এর বিবরণ বাদ দাও।
Type : নিচের Type Selection Method অনুসরণ করে একটি type নির্বাচন করুন।
Difficulty : 'High (Bar Exam MCQ)'।
Keywords : 3-5 legal terms or concepts central to the question, in Bangla.
Cited Acts and Sections : The primary Act name (in English) and the specific legislative address, or 'N/A — General Principle' where applicable.

[IRAC_Reasoning Keys — সকল মান বালিস]
Note: IRAC_Reasoning হলো Answer-এ পৌছানোর অভ্যন্তরীণ reasoning chain। মান-নিয়ন্ত্রণের জন্য JSON-এ সারকিক থাকবে, কিন্তু সকল UI/chatbot-এ লুকিয়ে রাখতে হবে।
- 'Issue' : এক বাক্য। বার পরীক্ষার প্রতিটি কোন মূল আইনি বিষয় পরীক্ষা করছে তা কখন। fill-in-the-blank বা MCQ কাঠামো বাদ দিয়ে শুধু মূল আইনি সমস্যাটি চিহ্নিত করুন।
- 'Rule' : প্রতিটি Possible Section-এর জন্য একটি করে key সহ একটি object। প্রতিটি গঠিত প্রতিটি section-এর জন্য ঠিক এক বাক্য সেই section-টি কী নিয়ন্ত্রণ করে তা প্যারামেজ করুন। কোনো section-এর পাঠ পুনর্লপন করবেন না। এক বাক্য, কোনো ব্যতিক্রম নেই।
- 'Application' : প্রতিটি Possible Section-এর জন্য একটি করে key সহ একটি object। গঠিত distractor section-এর প্রত্যেকটির জন্য ঠিক এক বাক্য স্পষ্টভাবে বাধ্য করুন কেন এটি এই বার পরীক্ষার প্রশ্নের উত্তর দিতে পারে না — scope mismatch, ভিন্ন trigger condition, বা বিষয়বস্তু পার্থক্যের জিহ্বিত। নিয়ন্ত্রণকারী ধারার জন্য এক বাক্য। ধারা কখন কখনো বিভাজ্য বা এটি ধারার সঠিক MCQ বিকল্পটি নিশ্চিত করে। শুধু দাবি নয়, আইনি কারণ দেখান। এক বাক্য, কোনো ব্যতিক্রম নেই। প্রতিটি entry-তে elimination ব্যাকের শুদ্ধ পঠন পরিবর্তন করুন — একই phrase বা বাক্য-পঠন দিয়ে প্রতিটি entry-র distractor rejection শুরু করবেন না।

বিশেষ ট্রিগার: Application-এর পরে reasoning chain সম্পূর্ণ। Answer হলো plain-language output। JSON-এ Conclusion key যোগ করবেন না।
[Answer Field — ব্যবহারকারীর জন্য একমাত্র দৃশ্যমান উত্তর]
Answer ফ্রেমটি IRAC_Reasoning-এর ঠিক পরে স্থাপন করতে হবে।

নিয়মাবলী:
- Act-এর নাম (ইংরেজিতে) দিয়ে শুরু করুন, তারপর সঠিক legislative address এবং 'অনুযায়ী'। যেমন:
Section: "The Evidence Act, 1872, Section 32, Clause (1) অনুযায়ী, ..."
Order and Rule: "The Code of Civil Procedure, 1908, Order XXXIX, Rule 1 অনুযায়ী, ..."
Schedule/Article: "The Limitation Act, 1908, Schedule I, Article 23 অনুযায়ী, ..."
Combined: Act-এ যে কখন আছে সেই ক্রমে কখন প্রযোজ্য কর (Subsection / Clause / Illustration / Rule) অন্তর্ভুক্ত করুন।
Section Number যদি 'General Principle' হয়, তাহলে 'সধারণ আইনি নীতি অনুযায়ী, ...' দিয়ে শুরু করুন এবং কোনো নির্দিষ্ট ধারা উল্লেখ করবেন না।
- 'সঠিক ধারা হলো...' বা এই ধরনের meta-language ব্যবহার করবেন না — সরাসরি প্রশ্নের উত্তর দিন।
- এক থেকে দুই মূল বাক্যে আইনি অবস্থান বাধ্য করুন, তারপর সঠিক বিকল্পের অক্ষর উল্লেখ করে একটি পৃথক বাক্য শেষ করুন এই নির্দিষ্ট ফর্ম্যাটে:
"সুতরাং সঠিক উত্তর হলো "(খ)"।" — প্রকৃত অক্ষরটি প্রতিস্থাপন করুন।
- মোট ৩৫-৬০ শব্দ (বিকল্প অক্ষর উপসংহার সহ)।
- উদ্ভূতির জন্য প্রয়োজনীয় পরিভাষার ব্যতীে পুনরাবৃত্তি করুন: "...নিশ্চিত করে যে "(ক)"।" সঠিক।"

উদাহরণ:
"Answer": "The Code of Civil Procedure, 1908, Section 11 অনুযায়ী, একটি আদালত এমন কোনো মামলা বিচার করতে পারে না যেখানে একই বিষয় একই পক্ষের মধ্যে ইতোমধ্যে সন্ধন আদালত কর্তৃক চূড়ান্তভাবে নিষ্পত্তি হয়েছে। — এটিই res judicata বিধির মূল বিষয়। সুতরাং সঠিক উত্তর হলো "(খ)"।"

[NO_IRAC_Reasoning Field — ব্যবহারকারীর জন্য নয়, training signal হিসেবে]
Answer-এর ঠিক পরে স্থাপন করতে হবে।

নিয়মাবলী:
- দুই থেকে তিন বাক্যে সহজ বাধ্য। Issue / Rule / Application / Conclusion পেন্সে ব্যবহার করা যাবে না।
- সম্পূর্ণ legislative citation অবশ্যই prose-এর মধ্যে স্বাভাবিকভাবে অন্তর্ভুক্ত করতে হবে — Act-এর নাম এবং legislative address। Section Number যদি 'General Principle' হয়, তাহলে সঠিক মতবাদসংক্রান্ত বিষয় উল্লেখ করুন।
- কেন সঠিক বিকল্পটি নিয়ন্ত্রণকারী ধারা দ্বারা সমর্থিত এবং সবচেয়ে কাছেই distractor বিকল্পটি কেন সমর্থিত নয় — উভয়ই সহজ ভাষায় বাধ্য করুন।
- Section Text ছব্ব পুনর্লপন করবেন না। শুধু প্যারামেজ করুন।
- শেষ বাক্যে সঠিক বিকল্পের অক্ষরটি বন্ধনীতে পুনরাবৃত্তি করুন: "...নিশ্চিত করে যে "(ক)"।" সঠিক।"
- লক্ষ্যমাত্রা: ৪৫-৬০ শব্দ।

উদাহরণ:
"NO_IRAC_Reasoning": "The Code of Civil Procedure, 1908, Section 11 একটি আদালতকে এমন মামলা বিচার করতে নিষেধ করে যেখানে একই বিষয় একই পক্ষের মধ্যে পূর্বে চূড়ান্তভাবে নিষ্পত্তি হয়েছে। অন্য বিকল্পগুলো যে section নথিগতভাবে উল্লেখ করে সেগুলো মামলা হয়েছে বন্ধ বা আদালতের সাধারণ ক্ষমতা নিয়ন্ত্রণ করে, res judicata নিষেধাজ্ঞা নয়, নিশ্চিত করে যে "(খ)"।" সঠিক।"

[Distractor Selection Logic]
নিয়ন্ত্রণকারী ধারার তো একই আইনি ডেটাসেট থেকে গঠিত distractor section নির্বাচন করতে হবে। নিম্নলিখিত প্রক্রিয়া অনুসরণ করুন:
Step 1 — নিয়ন্ত্রণকারী ধারার মূল আইনি ধারণাগুলো চিহ্নিত করুন: এটি কী নিয়ন্ত্রণ করে? কী মূল পরিভাষা ব্যবহার করে? কোন পদ্ধতিগত পর্যায়ে প্রযোজ্য?
Step 2 — আইনি ডেটাসেটের অন্য সকল section-কে নিম্নলিখিত মানদণ্ডে মূল্যায়ন করুন এবং সর্বোচ্চ সামগ্রিক মিল আছে এমন গঠি নির্বাচন করুন:
(ক) একই বা সম্পর্কিত আইনি বিষয়বস্তু
(খ) একই বা ওভারল্যাপিং মূল পরিভাষা
(গ) একই বা অনুরূপ পদ্ধতিগত প্রেক্ষাপটে প্রযোজ্য
(ঘ) প্রথম দুইতে সঙ্গত্য উত্তর মনে হয় কিন্তু সত্যি সত্যি বিলম্বে scope সীমাবদ্ধতা, ভিন্ন trigger condition বা সংকীর্ণ/বিস্তৃত প্রয়োগের কারণে বাতিল হয়
Step 3 — যাচাই করুন যে নির্বাচিত গঠি distractor-এর কোনোটিই বার পরীক্ষার প্রশ্নের উত্তর দেয় না, কিন্তু প্রতিটি সূক্ষ্ম বিশ্লেষণের আগে একজন যুক্তিসঙ্গত পাঠকের কাছে সঙ্গত্য উত্তর হিসেবে মনে হতে পারে।

Bangla bar-exam-style generation prompt, part 1 of 2.

Step 4 — Possible Sections-এ নিয়ন্ত্রণকারী ধারার অবস্থান randomly পরিবর্তন করুন। Generation batch জুড়ে প্রতিটি অবস্থানে (১ম থেকে ৫ম) প্রায় সমান ফ্রিকোয়েন্সিতে রাখুন।

Step 5 — যখন কোনো একক ধারা সত্যার প্রযোজ্য নয় (Section Number হলো "General Principle"), সমস্ত ৫টি Possible Section সম্ভাব্য প্রার্থী কিন্তু কোনোটিই সম্পূর্ণ নির্ণায়ক নয়। এই ক্ষেত্রে Application অবশ্যই সব ৫টি বাতিল করতে হবে এবং Answer-কে সম্ভাব্য আর্থনি নীতি থেকে যুক্তি দিতে হবে।

[Type Selection Method]
সহায়ক ফাইল bar_exam_question_type_taxonomy.mind দেখুন।
এই ফাইলে বাংলাদেশ বার কৌশলিক পরীক্ষার নির্দিষ্ট ১৫টি প্রশ্নের ধরন (T01–T15) সংজ্ঞায়িত এবং প্রতিটির জন্য উদাহরণ ও বিধিমালা দেওয়া আছে।

Type পরিচিতি (T01–T15):
T01 — Section / Provision Identification
T02 — Fill-in-the-Blank (Statutory Language)
T03 — Definition / Concept Identification
T04 — Scenario / Fact Pattern Application
T05 — Numerical / Quantitative
T06 — Negative Identification (Which is NOT)
T07 — Authority / Who Can (Power Identification)
T08 — Procedural Consequence / What Happens Next
T09 — Classification / How Many
T10 — Comparative / Which Has Higher
T11 — Professional Ethics Scenario
T12 — Remedy Identification
T13 — Order / Rule Identification
T14 — Timing / Stage of Proceedings
T15 — Correct / Incorrect Statement

প্রক্রিয়ার পরীক্ষার প্রশ্ন শ্রেণীবদ্ধ করার প্রক্রিয়া:
Step 1 — bar_exam_question_type_taxonomy.mind-এর T01–T15 সংজ্ঞার বিপরীতে question_text পড়ুন এবং শ্রেণিবদ্ধ করুন।
Step 2 — যে type-টি প্রদত্ত যে মাধ্যম ক্রিয়াকর্ম (cognitive operation) পরীক্ষা করে তা সবচেয়ে সঠিকভাবে বর্ণনা করে সেটি নির্বাচন করুন। বার পরীক্ষার প্রশ্নগুলোর বিষয়বস্তু নির্দিষ্ট — যে type মেনে না তা জোর করে বসাবেন না।
Step 3 — type-টি ID এবং নাম সহ লিখুন, যেমন "T04 — Scenario / Fact Pattern Application"
Step 4 — যে কোনো ১০টি entry-র যাতে কমপক্ষে ৬টি ভিন্ন type নিশ্চিত করুন এবং পরস্পর তিনবারের বেশি একই type না আসে তা যাচাই করুন।

[Constraint Guidelines]
Verbatim Integrity : Possible Sections-এর সমস্ত "Full Text" এবং "Section Text" আইনি ডেটাসেট থেকে হুবহু কপি করতে হবে। স্মৃতি থেকে পুনর্নির্মাণ বা সংক্ষিপ্ত করা যাবে না। যদি সঠিক পাঠ পাওয়া না যায়, entry-টি incomplete হিসেবে চিহ্নিত করুন।
Anti-Hallucination : স্মৃতি থেকে কখনো section-এর পাঠ পুনর্নির্মাণ বা প্যারাফ্রেজ করবেন না।
Verbatim Question : Question field অবশ্যই বার পরীক্ষার JSON-এর question_text-এর হুবহু কপি হতে হবে। fill-in-the-blank stem পরিবর্তন করা যাবে না।
Option-Letter Ending : সব Answer (Cancelled entry ছাড়া) অবশ্যই চিহ্ন একাটি বাক্যের শেষ হতে হবে যেখানে এই ফর্ম্যাটে সঠিক বিকল্পের অক্ষর থাকবে: "সুতরাং সঠিক উত্তর হলো "(ক)"।"
Correct Option Accuracy : Correct Option অবশ্যই বার পরীক্ষার JSON-এর correct_answer-এর সাথে হুবহু মিলতে হবে। বার পরীক্ষার answer key কখনো override করা যাবে না।
Structural Order : উপরে উল্লিখিত ক্রমে JSON keys অনুসরণ করুন।
IRAC Rule Step : শুধু প্যারাফ্রেজ করুন। IRAC Reasoning-এর ভেতরে section-এর পাঠ পুনরুৎপাদন করবেন না।
IRAC Application : Distractor elimination অবশ্যই আইনগতভাবে যুক্তিসংগত হতে হবে। যে distractor-টি নিয়ন্ত্রণকারী ধারার যত বেশি কাছাকাছি, তার elimination তত বেশি সুনির্দিষ্ট হতে হবে। শুধু দাবি ("এই ধারাটি প্রযোজ্য নয়") গ্রহণযোগ্য নয়।
Length Discipline : কোনো field তার শব্দসীমা অতিক্রম করলে আউটপুটের আগে সংশোধন করুন। সংক্ষিপ্ততা একটি মান-নিয়ন্ত্রণ প্রয়োজনীয়তা।
Answer Visibility Rule : IRAC Reasoning কোনো ব্যবহারকারীমুখী interface বা chatbot-এ প্রদর্শন করা যাবে না।
Cancelled Questions : Cancelled true হলো কোনো উত্তর তৈরি করা যাবে না। Answer এবং IRAC Reasoning Cancelled field নিয়মাবলী অনুযায়ী সেট করুন।
No Citation Artifacts : কোনো field-এ [cite: 1], [cite: 3], [cite: 1, 2] বা এরবনের কোনো citation marker ব্যবহার করা যাবে না। এগুলো RAG pipeline artifact — SFT output data-তে কোনো মতাবহ্য এগুলো থাকা উচিত নয়।
Entry Freshness : প্রতিটি entry ভিন্ন বার পরীক্ষার প্রশ্ন ব্যবহার করে — কোনো deduplication প্রয়োজন নেই, তবে IRAC Application opener প্রতিটি entry-তে ভিন্ন হওয়া নিশ্চিত করুন।

[Pre-Output Quality Check — আউটপুট দেওয়ার আগে প্রতিটি entry-র জন্য নিচের সকল বিষয় যাচাই করুন]
□ Question — বার পরীক্ষার JSON-এর question_text-এর হুবহু কপি
□ Bar Exam Options — বার পরীক্ষার JSON-এর options-এর হুবহু কপি
□ Correct Option — বার পরীক্ষার JSON-এর correct_answer-এর সাথে মিল
□ Possible Sections-এর সকল "Full Text" — আইনি ডেটাসেট থেকে হুবহু কপি, স্মৃতি থেকে পুনর্নির্মাণ নয়
□ Section Text — আইনি ডেটাসেট থেকে হুবহু কপি (অথবা নিয়মানুযায়ী "N/A")
□ Answer শেষ হয়েছে চিহ্ন এই ফর্ম্যাটে: "সুতরাং সঠিক উত্তর হলো "(ক)"।" সঠিক অক্ষর দিয়ে
□ Cancelled field — সঠিকভাবে সেট করা; Cancelled entry-তে কোনো তৈরি করা Answer নেই
□ Possible Sections-এ নিয়ন্ত্রণকারী ধারার অবস্থান — randomly পরিবর্তিত, পুনরাবৃত্তি গ্রহণযোগ্য নয়
□ IRAC Reasoning Rule এবং Application — প্রতিটি section-এর জন্য চিহ্ন এক বাক্য, কোনো ব্যতিক্রম নেই
□ Answer শব্দ সংখ্যা — ৩৫–৬০ শব্দের মধ্যে
□ NO IRAC Reasoning — প্রশ্নের সম্পূর্ণ উত্তর দিয়েছে, সঠিক ধারা উদ্ভূত, "সুতরাং সঠিক উত্তর হলো "(ক)"।" ফর্ম্যাটে শেষ হয়েছে
□ NO IRAC Reasoning — অন্য চিহ্ন উত্তরও কখনো কোন ধারা ভুল বলেনি
□ কোনো field-এ [cite: X] বা সম্পূর্ণ citation marker নেই

Figure 7: Bangla bar-exam-style generation prompt, part 2 of 2, reproduced from the original report.

G Original Error Exhibits

We inspected these traceable outputs without a sampling frame or coding protocol, so they illustrate failure modes rather than estimate prevalence. In a 2022 English item, BM25 supplied the Penal Code’s spousal exception to both 0.8B models. The base model repeated the provision but chose “siblings” in two of three runs; the fine-tuned model applied it and chose “husband/wife” in all three. A correct option still does not prove faithful use of the statute.

A 2022 Bangla item combines several failures. The fine-tuned 2B model gave a wrong answer with a fabricated subsection. The base model reached the correct option with faulty reasoning in English, and the parser captured a letter inside the fabricated citation rather than the final choice. Accuracy alone does not measure these citation, language, reasoning, and parsing failures, which is why the headline result remains bounded by expert review concerns (Aftahee et al., 2025).

Input Query	
Q: Harbours an offender is not punishable if the parties are: Options: A = Father/Son B = Guardian C = Husband/Wife ✓ D = Siblings	
Base Model Response (0.8B - BM25)	Fine-tuned Model Response (0.8B - BM25)
Answer: D (Siblings) X Wrong “...there is a specific exception provided in Section 216, which states that the provision does not extend to the case in which the harbour or concealment is by the husband or wife of the person to be apprehended... Therefore, the answer is (D).” Model correctly reads the husband/wife exception aloud — then ignores it and picks Siblings.	Answer: C (Husband/Wife) ✓ Correct “Under The Penal Code, 1860, section 212, Clause (b), harbouring a person is not punishable if that person is the husband or wife of the offender. This exception exists because the law explicitly excludes cases involving spouses. Therefore, (C) Husband/Wife is correct.” Applies the spousal exception correctly and directly.

(a) 0.8B, English item, BM25

Input Query	
প্রশ্ন: CPC 1908 অনুযায়ী 'Review' এর দরখাস্ত মঞ্জুর হলে তার বিরুদ্ধে প্রতিকার কী? বিকল্প: ক=রিভিশন খ=আপিল ✓ গ=রেফারেন্স ঘ=কোনো প্রতিকার নেই	
Base Model Response (2B - BM25)	Fine-tuned Model Response (2B - BM25)
Answer: খ (আপিল / Appeal) ✓ Correct “...A Review (Section ১১৪) is an appeal that can be taken against certain orders... the standard legal remedy for challenging such an order in the context of the CPC is to file an appeal. Therefore, the correct answer is filing an appeal.” Misframes s.114 (review ≠ appeal), but correctly concludes: remedy is appeal. It also answers the Bangla question in English.	Answer: ঘ (কোনো প্রতিকার নেই) X Wrong “The Code of Civil Procedure, 1908, Section ১১৪(গ) অনুযায়ী Review এর দরখাস্ত মঞ্জুর হলে তার বিরুদ্ধে আপিল করার কোনো সুযোগ নেই। সঠিক উত্তর (ঘ) কোনো প্রতিকার নেই।” Invents s.114(গ) — does not exist. Claims no appeal available — exact opposite of the law.

(b) 2B, Bangla item, BM25

Input Query	
Q: Under what circumstances can the Court order a Local Investigation under CPC 1908? Options: A = To clarify any doubtful point ✓ B = Property disputes only C = Contempt cases D = Criminal matters	
Base Model Response (4B - BM25)	Fine-tuned Model Response (4B - BM25)
Answer: B (Property disputes only) X Wrong “Based on the code of civil procedure, a local investigation can be ordered under circumstances involving property disputes, boundary disputes, or other matters in which a local investigation would be appropriate. These investigations are performed by the court commissioner appointed by the Court.” Restricts scope to property/boundary disputes — an invention. CPC Order 26 R.9 allows investigation to clarify any doubtful point.	Answer: A (To clarify any doubtful point) ✓ Correct “Under The Code of Civil Procedure, 1908, Order 26, Rule 9, the Court may order a local investigation to clarify any doubtful point. A commissioner is appointed to investigate the issue and report to the Court. Therefore, (A) is correct.” Accurately cites Order 26 R.9, correctly states the broad ‘any doubtful point’ scope, and identifies the commissioner mechanism.

(c) 4B, English item, BM25

Figure 8: Original error-analysis exhibits from the prior report, preserved as images. Exhibits (a) and (b) correspond to the illustrative cases in Section 7. Exhibit (c) is preserved as an artifact but is not used for a textual claim because its source run could not be traced.